Nativa, Cuiabá, v. 14, n. 4, e21893, 2026.
Pesquisas Agrárias e Ambientais
DOI: https://doi.org/10.31413/nat.v14i4.21893
ISSN: 2318-7670

Redes convolucionales y transformers para clasificar la pudrición negra del cacao: Auditoría de identidad de imágenes y sensibilidad al contexto de captura

Juan Carlos VALERO-GÓMEZ *[1], Juan Carlos CLARES-PERCA [1], Alex Peter ZÚÑIGA-INCALLA [1]

[1] Universidad Nacional de Moquegua, Moquegua, Perú.
*E-mail: jvalerog@unam.edu.pe

Enviado el: 04/07/2026; Aceptado el: 06/10/2026; Publicado el: 19/10/2026.

RESUMEN: La pudrición negra de la mazorca (BPR), causada por oomicetos del género *Phytophthora*, es una de las enfermedades más destructivas del cacao. En este trabajo se compararon cinco arquitecturas preentrenadas en ImageNet (Swin-T, ConvNeXt-T, DeiT-Ti, EfficientNetV2-S y ResNet50) para clasificar mazorcas sanas y con BPR en 2211 imágenes de un repositorio público. Antes del entrenamiento, se auditó la identidad de las imágenes mediante huellas perceptuales, descriptores DINOv2, correspondencias SIFT y revisión visual; las tomas de una misma mazorca o escena, presentes en el 79 % de las imágenes, se agruparon en 664 grupos. Los modelos se entrenaron con un protocolo común y se evaluaron mediante validación cruzada estratificada por grupos (tres repeticiones de cinco pliegues), pruebas de DeLong y McNemar, y la corrección de Holm. El AUC medio varió entre 0.989 y 0.993 y la exactitud entre 0.954 y 0.964. Swin-T tuvo la mayor sensibilidad y DeiT-Ti el menor costo computacional, pero las diferencias en AUC solo fueron significativas en algunas comparaciones y repeticiones. En 1958 imágenes sanas no utilizadas, la especificidad (0.952 a 0.965) fue similar a la de la validación cruzada, excepto en una escena de captura que, en el entrenamiento, solo aparecía en mazorcas enfermas, donde bajó a valores de 0.52 a 0.85. Las cinco arquitecturas discriminan bien y de forma similar, pero también aprenden asociaciones contextuales, por lo que su uso en campo exige una validación externa con imágenes de ambas clases tomadas en los mismos contextos.
Palabras clave: *Phytophthora* spp.; aprendizaje profundo; transferencia de aprendizaje; redes neuronales convolucionales; validación cruzada por grupos; Oomicetos.

Convolutional networks and transformers for classifying cacao black pod rot: image identity audit and sensitivity to capture context

ABSTRACT: Black pod rot (BPR), caused by oomycetes of the genus *Phytophthora*, is one of the most destructive diseases of cacao. We compared five ImageNet-pretrained architectures (Swin-T, ConvNeXt-T, DeiT-Ti, EfficientNetV2-S, and ResNet50) for classifying healthy and BPR-affected pods in 2,211 images from a public repository. Before training, image identity was audited with perceptual hashes, DINOv2 descriptors, SIFT correspondences, and visual review; shots of the same pod or scene, present in 79% of the images, were merged into 664 groups. The models were trained with a common protocol and evaluated with group-stratified cross-validation (three repetitions of five folds), DeLong and McNemar tests, and Holm correction. Mean AUC ranged from 0.989 to 0.993 and accuracy from 0.954 to 0.964. Swin-T had the highest sensitivity and DeiT-Ti the lowest computational cost, but AUC differences were significant only in some comparisons and repetitions. On 1,958 unused healthy images, specificity (0.952 to 0.965) was similar to that of cross-validation, except for a capture scene that appeared only in diseased pods during training, where it dropped to 0.52 to 0.85. The five architectures discriminate well and similarly, but they also learn context associations, so field use requires external validation with images of both classes taken in the same contexts.
Keywords: *Phytophthora* spp.; deep learning; transfer learning; convolutional neural networks; group cross-validation; Oomycetes.

1. INTRODUCCIÓN

El cacao (*Theobroma cacao*) es un cultivo estratégico para la economía agrícola mundial y de suma importancia para el desarrollo socioeconómico de Latinoamérica y de países exportadores como el Perú, donde miles de familias dependen de su comercialización. La producción, sin embargo, convive con enfermedades que reducen el rendimiento de forma recurrente. Entre ellas, la pudrición negra de la mazorca (BPR, por sus siglas en inglés) es la de mayor distribución en el mundo (MARELLI et al., 2019). La causan varias especies de *Phytophthora*, que no son hongos sino oomicetos: organismos del filo Oomycota, dentro del reino Stramenopila (Chromista), más cercanos filogenéticamente a las algas heterocontas que a los hongos

verdaderos, con paredes celulares compuestas principalmente por β-glucanos y celulosa en lugar de quitina (BEAKES et al., 2012; KAMOUN et al., 2015). Esta distinción no es solo taxonómica, ya que varios fungicidas de uso común tienen poca eficacia contra los oomicetos. *P. palmivora*, la especie de distribución más amplia, provoca pérdidas de entre 20 y 30 % de las mazorcas y puede matar hasta 10 % de los árboles al año debido a cancros en el tallo; *P. megakarya* predomina en África occidental, mientras que *P. capsici* y *P. citrophthora* se registran en Centro y Sudamérica (GUEST, 2007; PERRINE-WALKER, 2020).

La vigilancia de la BPR se basa todavía en la inspección visual de las mazorcas, una tarea lenta que depende de la experiencia del evaluador y produce estimaciones inconsistentes de la severidad (TAN et al., 2018). El aprendizaje profundo ha cambiado este panorama en la fitopatología en general: desde los primeros trabajos sobre hojas fotografiadas en laboratorio (Mohanty et al., 2016) hasta revisiones que documentan cientos de estudios con redes neuronales convolucionales (CNN) (ABADE et al., 2021). En cacao, Yanac-Montesino et al. (2021) aplicaron ResNet18 a la detección de *P. palmivora* con una exactitud cercana al 83 %; Godmalin et al. (2023) clasificaron niveles de infección mediante transferencia de aprendizaje y trabajos más recientes reportan exactitudes o valores de mAP de entre 94 y 99 % con CNN propias, EfficientNet o detectores YOLO (PRADANA et al., 2024; ROLA et al., 2024; SUDJUD et al., 2025; MAYLIANTI et al., 2025; TECHIE-MENSON et al., 2026).

En paralelo, la visión computacional incorporó dos familias de arquitecturas que aún se han explorado poco en el contexto del cacao. Los Vision Transformers aplican autoatención a parches de la imagen; DeiT mostró que pueden entrenarse con conjuntos de datos del tamaño de ImageNet (Touvron et al., 2021) y Swin Transformer restringió la atención a ventanas desplazadas para obtener representaciones jerárquicas (LIU et al., 2021). Por otro lado, las CNN se modernizaron: EfficientNetV2 optimizó conjuntamente la precisión y la velocidad de entrenamiento (Tan; Le, 2021) y ConvNeXt incorporó en una red convolucional varias decisiones de diseño de los transformers (LIU et al., 2022). Navarro-Cabrera et al. (2026) compararon ResNet50, EfficientNet-B0 y ViT-B/16 en imágenes de cacao y no encontraron diferencias significativas entre ellas, lo que sugiere que el protocolo experimental pesa tanto como la arquitectura.

Esa última observación apunta a un problema metodológico en la literatura. Una buena parte de los estudios sobre cacao reporta una única partición de entrenamiento y prueba, sin intervalos de confianza ni pruebas de significancia, y rara vez examina si el conjunto de datos contiene fotografías repetidas o atajos que inflen las métricas (BARBEDO, 2022; KAPOOR; NARAYANAN, 2023). Con muestras de prueba de unos cientos de imágenes, las diferencias de uno o dos puntos porcentuales en la exactitud entre arquitecturas pueden deberse únicamente a la partición de los datos.

El objetivo de este estudio fue comparar cuatro arquitecturas modernas y una referencia clásica, Swin-T, ConvNeXt-T, DeiT-Ti, EfficientNetV2-S y ResNet50, en la clasificación de mazorcas sanas y con BPR bajo un protocolo común y controlado: auditoría de la identidad de las imágenes para agrupar las fotografías de una misma mazorca o escena, validación cruzada repetida por grupos, pruebas estadísticas pareadas, análisis de calibración, mapas de activación, medición del costo computacional y evaluación de la especificidad en imágenes sanas no utilizadas, desglosada por escena de captura. Además de ordenar las arquitecturas, interesaba saber hasta qué punto las métricas dependen de que las imágenes de entrenamiento y de prueba sean realmente independientes, y en qué condiciones de captura fallan los modelos.

2. MATERIAL Y MÉTODOS

2.1. Conjunto de datos

Se utilizó el repositorio público *Cacao Diseases* (Pagaduan, 2021), que reúne fotografías de mazorcas tomadas en plantaciones, con fondos naturales de follaje, troncos y hojarasca y con iluminación variable. El repositorio incluye tres clases: mazorcas sanas, mazorcas con BPR y mazorcas dañadas por el barrenador (*pod borer*). La clase barrenador se excluyó porque el estudio se planteó como una clasificación binaria centrada en la BPR y porque sus 103 imágenes, muy pocas frente a las otras dos clases, solo habrían permitido estimaciones imprecisas. La clase sana es mucho más numerosa en el repositorio (3344 imágenes) que la clase con BPR (943); para acercar ambas clases a un tamaño similar, se tomaron 1268 imágenes sanas, correspondientes a tramos de numeración consecutiva del repositorio, elegidos sin un criterio relacionado con el contenido de las imágenes, y se conservaron todas las imágenes con BPR. No se conservó un registro del procedimiento de sorteo de esos tramos; por ello, la lista de los archivos empleados se publica junto con el código, de modo que el subconjunto puede reproducirse exactamente.

El conjunto analizado estuvo compuesto por 2211 imágenes en color: 1268 de mazorcas sanas (57.3 %) y 943 con síntomas de BPR (42.7 %), caracterizados por lesiones necróticas de color parduzco a negro que avanzan sobre el epicarpio (Figura 1). Las etiquetas son las del repositorio original; no fueron revisadas por un fitopatólogo y el repositorio no identifica la especie de *Phytophthora* involucrada, por lo que en todo el texto se emplea la denominación *Phytophthora* spp.

Las imágenes tienen dos resoluciones, que se trataron como estratos y no como fuentes de captura confirmadas, porque el repositorio no informa la cámara, la finca ni la fecha. El estrato A reúne 1829 imágenes de 1080 × 1080 píxeles (900 sanas y 929 con BPR) y el estrato B, 382 imágenes de 2160 × 2160 píxeles (368 sanas y 14 con BPR). Este desbalance se tuvo en cuenta en el control de atajos descrito en la Sección 2.2.

2.2. Auditoría de identidad y control de atajos

Antes de formar las particiones se buscaron fotografías de una misma mazorca o de una misma escena, porque su presencia simultánea en entrenamiento y en prueba infla las métricas (KAPOOR; NARAYANAN, 2023). El repositorio contiene, para muchas mazorcas, varias tomas sucesivas desde ángulos o distancias distintos, y esas tomas reciben números de archivo consecutivos. La auditoría combinó cuatro fuentes de evidencia para cada par de imágenes: la huella MD5 (duplicados exactos), la distancia de Hamming entre huellas perceptuales pHash de 64 bits, la similitud coseno entre descriptores del modelo autosupervisado DINOv2 ViT-S/14 (Oquab et al., 2024) y el número de correspondencias geométricamente consistentes entre puntos característicos SIFT (Lowe, 2004), estimado con una

homografía ajustada por RANSAC (Fischler; Bolles, 1981) sobre las imágenes reducidas a 640 × 640 píxeles. Se evaluaron como candidatos todos los pares de la misma clase con números de archivo a una distancia de 30 o menos, los diez vecinos más parecidos de cada imagen según DINOv2 y todos los pares con similitud coseno de 0.85 o más.

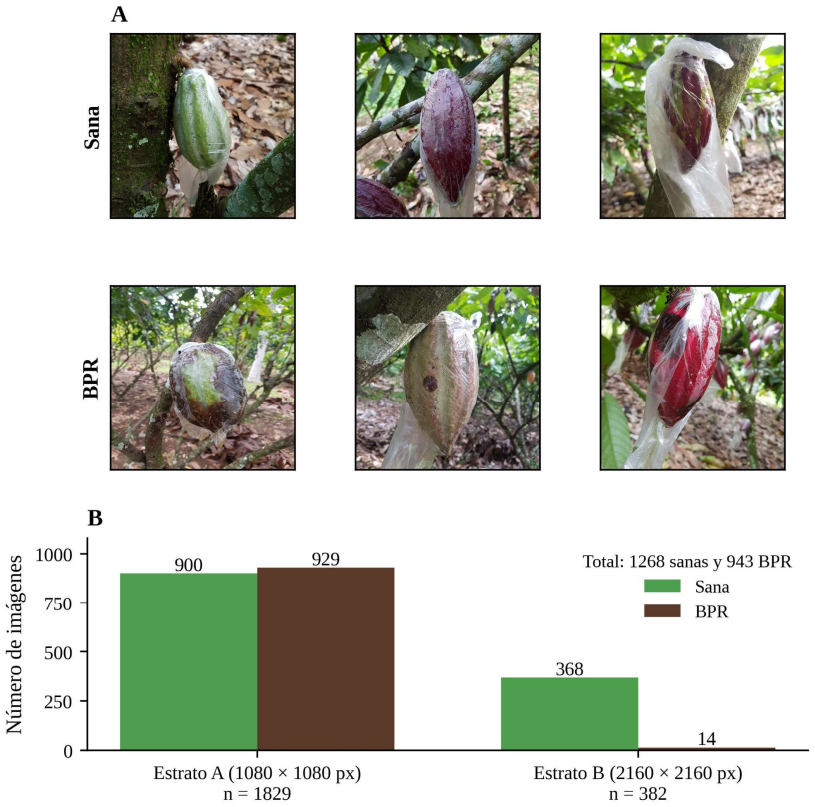


Figura 1. Ejemplos de mazorcas de cacao sanas y con pudrición negra (BPR) causada por Phytophthora spp. (A) y número de imágenes por clase y por estrato de resolución (B). El estrato A corresponde a imágenes de 1080 × 1080 píxeles y el estrato B a imágenes de 2160 × 2160 píxeles.

Figure 1. Examples of healthy cacao pods and pods with black pod rot (BPR) caused by Phytophthora spp. (A), and the number of images per class and resolution stratum (B). Stratum A corresponds to 1080 × 1080 pixel images and stratum B to 2160 × 2160 pixel images.

Dos imágenes se asignaron al mismo grupo cuando cumplían al menos una de estas condiciones: distancia pHash ≤ 6; similitud coseno ≥ 0.93; números de archivo a una distancia de 3 o menos con similitud coseno ≥ 0.75, lo que reúne las tomas sucesivas de una misma mazorca; o 40 o más correspondencias SIFT, que identifican la misma escena aunque cambie la mazorca fotografiada. Los umbrales se fijaron tras inspeccionar muestras de pares en cada franja de evidencia; en pares de escenas distintas elegidos al azar, el número de correspondencias no superó 11. Los 650 pares no unidos por estas condiciones, pero con evidencia intermedia (números a distancia de 4 a 10 con similitud ≥ 0.80, similitud entre 0.88 y 0.93, o entre 20 y 39 correspondencias), se presentaron en orden aleatorio para una revisión visual, en la que solo se mostraban las dos fotografías, sin nombres de archivo, clases ni predicciones. Se revisaron 396 de esos pares; los confirmados como la misma mazorca o escena se unieron, y los 254 no revisados se trataron como distintos. Los grupos resultantes nunca se repartieron entre los conjuntos de entrenamiento, validación y prueba.

Para detectar atajos, es decir, rasgos ajenos a la lesión que permitan acertar la clase (Geirhos et al., 2020), se entrenaron modelos triviales de regresión logística con cuatro grupos de variables: el tamaño original de la imagen, el color del borde de la imagen como aproximación al fondo, el color global (medias, desviaciones e histogramas en HSV) y la combinación de todas ellas. Estos modelos se evaluaron con los mismos pliegues que las redes. Su capacidad predictiva indica asociaciones entre el contexto visual y las etiquetas, pero superar estos controles simples no basta para descartar que las redes recurran a atajos más complejos.

2.3. Particiones y validación cruzada

El desempeño se estimó mediante validación cruzada estratificada por grupos (StratifiedGroupKFold) de cinco pliegues, repetida tres veces con semillas distintas (42, 142 y 242), lo que da un total de 15 ejecuciones por modelo. En cada pliegue, el conjunto de prueba se utilizó únicamente para la evaluación final. Del resto se separó una séptima parte (1/7), también estratificada por grupos, para la validación interna de la detención temprana, la reducción de la tasa de aprendizaje y la elección del umbral. Como cada grupo se asigna por completo a un solo conjunto y los grupos difieren mucho en tamaño, el número de imágenes y la proporción de clases variaron entre los pliegues (Tabla 1). Por esa razón, los resultados principales se calcularon sobre las predicciones fuera de pliegue: como cada imagen aparece exactamente una vez en prueba dentro de cada repetición, cada modelo produce una predicción por imagen, lo que permite estimaciones sobre las 2211 imágenes y comparaciones pareadas entre modelos.

2.4. Arquitecturas y protocolo de entrenamiento

Se evaluaron cinco arquitecturas de la biblioteca timm (Wightman, 2019) con pesos preentrenados en ImageNet-1k (Deng et al., 2009): Swin-T (Liu et al., 2021), ConvNeXt-T

(Liu et al., 2022), DeiT-Ti (Touvron et al., 2021), EfficientNetV2-S (Tan; Le, 2021) y ResNet50 (He et al., 2016), siendo esta última la referencia convolucional clásica. En todas se reemplazó el clasificador original por la misma cabecera: agrupación global propia de cada arquitectura, dropout de 0.5 y una capa lineal con una sola salida, inicializada con una distribución normal truncada con desviación estándar de 0.02 y sesgo nulo. La salida z de esa capa (logit) se transforma en la probabilidad de BPR mediante la función sigmoide (Ecuación 1).

$$\hat{p} = \sigma(z) = \frac{1}{1+e^{-z}} \qquad (01)$$

En la Ecuación 1, z es el logit que produce la red para una imagen y $\hat{p}$ la probabilidad estimada de que la mazorca presente BPR. Una imagen se asigna a la clase BPR cuando $\hat{p} \geq 0.5$. La función de pérdida fue la entropía cruzada binaria calculada sobre la salida del logit (Ecuación 2).

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log \hat{p}_i + (1-y_i)\log(1-\hat{p}_i)] \qquad (02)$$

En la Ecuación 2, N es el número de imágenes del lote, y_i es la etiqueta real (1 para BPR y 0 para sana) y $\hat{p}i$ es la probabilidad predicha para la imagen i. La pérdida castiga con fuerza las predicciones seguras pero erróneas (GOODFELLOW et al., 2016).

El entrenamiento siguió la estrategia de transferencia de aprendizaje en dos fases (Yosinski et al., 2014; Howard; Ruder, 2018; Kornblith et al., 2019), resumida en la Figura 2.

Tabla 1. Número de imágenes por clase en los conjuntos de entrenamiento, de validación interna y de prueba de cada pliegue (repetición 1).
Table 1. Number of images per class in the training, internal validation and test sets of each fold (repetition 1).

Pliegue	Entrenamiento: sana	Entrenamiento: BPR	Validación: sana	Validación: BPR	Prueba: sana	Prueba: BPR
1	1037	789	125	66	106	88
2	503	554	419	118	346	271
3	1092	677	101	96	75	170
4	662	629	129	82	477	232
5	501	595	503	166	264	182

Nota. Los grupos por secuencia de captura (664) se asignaron íntegramente a un único conjunto, por lo que los tamaños de los conjuntos difieren entre pliegues: en las tres repeticiones, los conjuntos de prueba contaron con entre 194 y 843 imágenes, y la proporción de BPR por conjunto varió entre 18.9 y 78.0 %. Las particiones completas están incluidas en el repositorio.
Note. Capture sequence groups (664) were assigned entirely to a single set; consequently, set sizes differed across folds: across the three replicates, test sets ranged from 194 to 843 images, and the proportion of BPRs per set varied between 18.9% and 78.0%. The complete partitions are included in the repository.

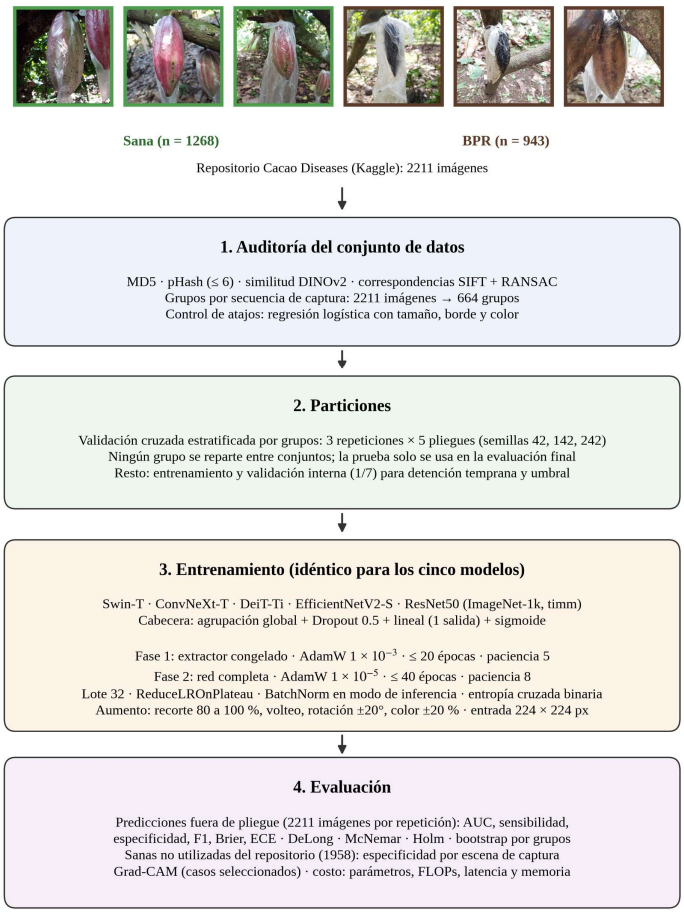


Figura 2. Flujo metodológico del estudio: auditoría del conjunto de datos y agrupación por secuencia de captura; particiones con validación cruzada repetida por grupos; entrenamiento en dos fases con cabecera común; y evaluación estadística, de calibración, de interpretabilidad y de costo computacional.
Figure 2. Methodological workflow: dataset audit and capture-sequence grouping, repeated group cross-validation splits, two-stage training with a common head and statistical, calibration, interpretability and computational cost evaluation.

En la primera fase se congeló el extractor de características y solo se entrenó la cabecera durante un máximo de 20 épocas con AdamW (Loshchilov; Hutter, 2019), tasa de aprendizaje de 1×10^{-3} y decaimiento de pesos de 0.01. En la segunda fase se descongeló la red completa, partiendo de los mejores pesos de la primera fase, y se entrenó durante un máximo de 40 épocas con una tasa de aprendizaje de 1×10^{-5}. En ambas fases se usaron lotes de 32 imágenes; se redujo la tasa de aprendizaje a la mitad tras tres épocas sin mejora en la pérdida de validación (ReduceLROnPlateau) y se aplicó detención temprana con paciencia de 5 y 8 épocas, respectivamente, restaurando siempre los pesos con la menor pérdida de validación. Las capas de normalización por lotes (Ioffe; Szegedy, 2015) de ResNet50 y EfficientNetV2-S se mantuvieron en modo de inferencia durante ambas fases, con sus parámetros afines entrenables en la segunda; así se evitó que las estadísticas de lotes pequeños alteraran bruscamente las características recibidas por la cabecera al descongelar la red.

Cada imagen se convirtió a RGB y se guardó una copia de 256 × 256 píxeles mediante interpolación bicúbica, que luego se redimensionó a 224 × 224 píxeles y se normalizó con la media y la desviación estándar de ImageNet correspondientes a cada modelo. Durante el entrenamiento se aplicó un aumento de datos moderado: recorte aleatorio de 80 a 100 % del área, volteo horizontal, rotación de hasta ±20° y variaciones de brillo, contraste y saturación de hasta ±20 %. El umbral alternativo se eligió en la validación interna: entre los umbrales que maximizan el F1 se tomó el punto medio del intervalo más amplio, lo que evita la inestabilidad del primer máximo cuando las clases están casi separadas. La Tabla 2 resume los hiperparámetros.

2.5. Entorno computacional

Los experimentos se implementaron en Python 3.10.21 con PyTorch 2.14.0 (Paszke et al., 2019) compilado para CUDA 12.6 y cuDNN 9.10.2, timm 1.0.30, torchvision 0.29.0, scikit-learn 1.7.2, SciPy 1.15.3, statsmodels 0.15.0 y pytorch-grad-cam 1.5.7. El entrenamiento se ejecutó en un servidor con dos GPU NVIDIA A100 de 80 GB (controlador 560.35.03), un procesador Intel Xeon Gold 5420+ de 112 hilos y 251 GB de memoria RAM, dentro de un contenedor Docker. Se entrenó con precisión mixta bfloat16 y la predicción final se calculó en precisión simple (float32), con el formato TF32 desactivado. Para aprovechar las GPU, se ejecutaron seis trabajos simultáneos por GPU mediante una cola dinámica; cada trabajo fijó sus propias semillas (42 + 1000r + k, donde r es la repetición y k el pliegue) y usó cuDNN en modo determinista.

Tabla 2. Hiperparámetros y configuración de entrenamiento, comunes a las cinco arquitecturas.
Table 2. Hyperparameters and training configuration, common to the five architectures.

Elemento	Configuración
Modelos (timm)	swin_tiny_patch4_window7_224.ms_in1k; convnext_tiny.fb_in1k; deit_tiny_patch16_224.fb_in1k; tf_efficientnetv2_s.in1k; resnet50.tv_in1k
Pesos iniciales	ImageNet-1k
Entrada	Copia RGB de 256 × 256 píxeles (interpolación bicúbica) redimensionada a 224 × 224; normalización con la media y la desviación de ImageNet de cada modelo
Cabecera	Agrupación global + Dropout (0.5) + capa lineal de 1 neurona (sin capa oculta) + sigmoide
Inicialización de la capa de salida	Normal truncada (desviación 0.02), sesgo 0
Función de pérdida	Entropía cruzada binaria sobre el logit
Optimizador	AdamW, decaimiento de pesos 0.01
Fase 1 (extractor congelado)	Tasa 1×10^{-3}; máximo 20 épocas; paciencia 5; lote 32
Fase 2 (red completa)	Tasa 1×10^{-5}; máximo 40 épocas; paciencia 8; lote 32
Planificador	ReduceLROnPlateau (factor 0.5, paciencia 3, tasa mínima 1×10^{-7})
Normalización por lotes	Modo de inferencia en ambas fases (ResNet50 y EfficientNetV2-S)
Aumento de datos	Recorte aleatorio 80 a 100 %; volteo horizontal; rotación ±20°; brillo, contraste y saturación ±20 %
Precisión numérica	bfloat16 en entrenamiento; float32 sin TF32 en la predicción final
Épocas ejecutadas (fase 1 / fase 2, media)	Swin-T 19.0 / 20.9; ConvNeXt-T 19.6 / 30.7; DeiT-Ti 16.7 / 19.7; EfficientNetV2-S 19.3 / 18.7; ResNet50 14.6 / 17.6
Tiempo por ejecución (min, media ± DE)	Swin-T 3.4 ± 0.6; ConvNeXt-T 3.8 ± 0.8; DeiT-Ti 1.4 ± 0.3; EfficientNetV2-S 3.2 ± 0.7; ResNet50 2.0 ± 0.5
Semillas	Particiones: 42, 142 y 242; entrenamiento: 42 + 1000r + k
Hardware y software	2 × NVIDIA A100 80 GB; PyTorch 2.14.0 (CUDA 12.6); timm 1.0.30

Nota. Los tiempos corresponden a 12 trabajos ejecutados en paralelo (6 por GPU). r: repetición; k: pliegue.
Note. The times correspond to 12 jobs executed in parallel (six per GPU). r: repetition; k: fold.

2.6. Métricas de evaluación

Las métricas se calcularon tomando BPR como clase positiva, tanto en el conjunto de prueba de cada pliegue como en las predicciones fuera de pliegue de cada repetición. A partir de los verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP) y falsos negativos (FN) se obtuvieron la exactitud, la precisión, la sensibilidad y la especificidad (Fawcett, 2006; Luque et al., 2019), definidas en las Ecuaciones 3 a 6.

$$\text{Exactitud} = \frac{VP+VN}{VP+VN+FP+FN} \quad (03)$$

$$\text{Precisión} = \frac{VP}{VP+FP} \quad (04)$$

$$\text{Sensibilidad} = \frac{VP}{VP+FN} \quad (05)$$

$$\text{Especificidad} = \frac{VN}{VN+FP} \quad (06)$$

La sensibilidad es la métrica de mayor interés agronómico, porque cada mazorca enferma no detectada sigue produciendo inóculo en la plantación. El F1 combina precisión y sensibilidad mediante su media armónica (Ecuación 7).

$$F_1 = \frac{2 \cdot \text{Precisión} \cdot \text{Sensibilidad}}{\text{Precisión} + \text{Sensibilidad}} \quad (07)$$

El F1 macro promedia los F1 de ambas clases. Como métrica principal de discriminación, se usó el área bajo la curva ROC (AUC), que no depende del umbral y equivale a la probabilidad de que una mazorca con BPR tenga una probabilidad mayor que la de una sana (Ecuación 8).

$$\text{AUC} = \frac{1}{n_1 n_0} \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} \psi\left(\hat{p}_i, \hat{p}_j\right) \quad (08)$$

En la Ecuación 8, n_1 y n_0 son los números de imágenes con BPR y sanas, respectivamente, y la función ψ vale 1 cuando la mazorca enferma i tiene mayor probabilidad que la sana j, 0.5 en caso de empate y 0 en caso contrario. También se reportó la precisión promedio (AP) en el espacio precisión-sensibilidad. La calibración de las probabilidades se evaluó mediante el puntaje de Brier (Brier, 1950) y el error de calibración esperado (ECE) con diez intervalos de igual anchura (Guo et al., 2017), definidos en las Ecuaciones 9 y 10. El puntaje de Brier combina discriminación y calibración, y ninguno de los dos índices garantiza por sí solo la calibración en una población con una prevalencia de la enfermedad distinta.

$$\text{Brier} = \frac{1}{N} \sum_{i=1}^{N} (\hat{p}_i - y_i)^2 \quad (09)$$

$$\text{ECE} = \sum_{b=1}^{B} \frac{n_b}{N} |\bar{y}_b - \bar{p}_b| \quad (10)$$

En la Ecuación 10, las predicciones se agrupan en B = 10 intervalos de probabilidad; n_b es el número de imágenes del intervalo b, $\bar{y}_b$ la proporción observada de BPR y $\bar{p}_b$ la probabilidad media predicha. Un modelo perfectamente calibrado tiene ECE igual a cero.

2.7. Análisis estadístico

Las métricas por pliegue se resumen como media ± desviación estándar de las 15 ejecuciones; esa desviación describe la variación entre ejecuciones dependientes y no equivale a un intervalo de confianza. Para comparar modelos se usaron las predicciones fuera de pliegue de cada repetición. El AUC de cada modelo y su intervalo de confianza del 95 % (IC95 %) se estimaron con el método de DeLong, y las diferencias entre pares de modelos se contrastaron con la prueba de DeLong para curvas ROC correlacionadas (DeLong et al., 1988), según la implementación rápida de Sun; Xu (2014). El estadístico se obtiene mediante la Ecuación 11.

$$z = \frac{\text{AUC}_A - \text{AUC}_B}{\sqrt{V_A + V_B - 2\,C_{AB}}} \quad (11)$$

En la Ecuación 11, V_A y V_B son las varianzas de los AUC de los modelos A y B, y C_{AB} su covarianza. Las tres se calculan a partir de los componentes estructurales de DeLong sobre las mismas imágenes, lo que tiene en cuenta que ambos modelos se evalúan con los mismos datos. La exactitud con umbral de 0.5 se comparó con la prueba de McNemar (McNemar, 1947; Dietterich, 1998), que solo considera las imágenes en las que los modelos discrepan (Ecuación 12).

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} \quad (12)$$

En la Ecuación 12, b es el número de imágenes que solo acierta el modelo A y c, el número que solo acierta el modelo B; cuando b + c < 25, se usó la versión exacta binomial. En ambas pruebas, se corrigieron los valores p de las diez comparaciones por pares mediante el procedimiento de Holm (Holm, 1979) en cada repetición. Los IC95 % de la exactitud y del F1 se estimaron mediante 2000 remuestreos bootstrap por grupos, en los que las fotografías de una misma secuencia de captura se remuestrean juntas. Estos contrastes son aproximados: las predicciones fuera de pliegue provienen de modelos entrenados con conjuntos de datos solapados, por lo que no capturan toda la variabilidad del reentrenamiento (BATES et al., 2024). Por la misma razón, no se aplicaron pruebas que traten las 15 ejecuciones como independientes, y la falta de significación no se interpretó como equivalencia entre los modelos.

2.8. Interpretabilidad y costo computacional

Como análisis cualitativo de casos, se generaron mapas de activación Grad-CAM (Selvaraju et al., 2017) con los modelos del primer pliegue de la primera repetición para tres mazorcas sanas y tres con BPR, elegidas al azar del conjunto de prueba, así como para las imágenes que Swin-T clasificó incorrectamente en ese pliegue. Cada mapa muestra la evidencia a favor de la clase predicha, utilizando el logit con signo negativo para la clase sana. Estos mapas no demuestran que una región cause la decisión (ADEBAYO et al., 2018). Las capas objetivo fueron el último bloque de ResNet50, la última etapa de ConvNeXt-T, la convolución final de EfficientNetV2-S y la normalización del último bloque de Swin-T y de DeiT-Ti, que reorganizan los tokens en su malla espacial. El costo computacional se describió mediante el número de parámetros, las operaciones de punto flotante (FLOPs) para una imagen de 224 × 224 píxeles, la latencia por imagen en GPU (float32 y bfloat16) y en CPU con cuatro hilos, el rendimiento en imágenes por segundo con lotes de 64 y la memoria máxima asignada en GPU. Cada modelo se midió en un proceso independiente, con las GPU libres y tres repeticiones, usando tensores sintéticos; por tanto, las cifras no incluyen la lectura, la decodificación ni el preprocesamiento de las imágenes.

2.9. Evaluación complementaria con imágenes sanas no utilizadas

Las 2076 imágenes sanas del repositorio que no formaron parte del conjunto analizado se reservaron como conjunto adicional de evaluación. Ninguno de los 75 modelos entrenados (cinco arquitecturas, tres repeticiones y cinco pliegues) tuvo acceso a ellas durante el entrenamiento, la validación interna o la elección del umbral, y para esta evaluación no se reentrenó ningún modelo. Las imágenes se identificaron mediante su huella MD5 y se procesaron de la misma manera que el conjunto principal. Se excluyeron las que la regla de la Sección 2.2 vinculaba con alguna imagen del conjunto principal, y, en un análisis de sensibilidad, se excluyeron además las que solo presentaban evidencia intermedia de vínculo. Como control, cada modelo volvió a

predecir su propio conjunto de prueba y se verificó que las probabilidades coincidieran con las obtenidas en la validación cruzada.

Como este conjunto contiene solo mazorcas sanas, permite estimar la especificidad (Ecuación 6) y el número de falsos positivos en imágenes nuevas. Como medida complementaria de discriminación se calculó, para cada modelo, un AUC con negativos nuevos (Ecuación 8), en el que los positivos son las mazorcas con BPR de su propio pliegue de prueba, ya empleadas en la evaluación principal, y los negativos son las sanas no utilizadas. La especificidad en estas imágenes, promediada sobre los cinco modelos de cada repetición, se comparó con la obtenida fuera de pliegue en las 1268 sanas del conjunto principal, con un IC95 % calculado mediante 2000 remuestreos bootstrap por grupos.

La especificidad se desglosó, además, por escena de captura. Durante la auditoría se observó una composición particular, la mazorca sostenida en la mano sobre un recipiente lleno de mazorcas, que en el conjunto principal aparecía solo en imágenes con BPR. Para identificarla sin recurrir a las predicciones, se calculó el centroide de los descriptores DINOv2 de esas imágenes y se asignó a la escena toda imagen con una similitud coseno de 0.80 o más respecto de ese centroide. El censo se aplicó por igual al conjunto principal y a las sanas no utilizadas.

3. RESULTADOS

3.1. Calidad del conjunto de datos

No se encontraron duplicados exactos. La distancia pHash de cada imagen a su vecina más cercana tuvo una mediana de 16, frente a 30 entre pares al azar, y solo un par (healthy_1572 y healthy_2755) quedó por debajo del umbral de 6; la similitud DINOv2 con la vecina más cercana tuvo una mediana de 0.86.

Las tomas sucesivas de una misma mazorca, en cambio, fueron frecuentes. Con la regla de la Sección 2.2 y la revisión visual, el conjunto quedó dividido en 664 grupos; los criterios de pHash y de similitud ≥ 0.95, aplicados por separado, habrían dado 2205 grupos. En total, 1750 imágenes (79 %) comparten grupo con al menos otra; la mediana de esos grupos es de 3 imágenes y 17 grupos superan las 20. El grupo mayor reúne 351 imágenes (309 sanas y 42 con BPR) de escenas encadenadas por similitud, y cuatro grupos, entre ellos el mayor, combinan ambas clases, en general porque incluyen mazorcas sanas y enfermas del mismo árbol. En la revisión visual, 68 de los 396 pares revisados (17.2 %) correspondieron a la misma mazorca o escena; si esa proporción se mantuviera en los 254 pares no revisados, quedarían unos 44 pares de este tipo sin agrupar.

En el control de atajos, la regresión logística alcanzó un AUC de 0.539 con el tamaño de la imagen, 0.683 con el color del borde, 0.776 con el color global y 0.863 con todas las variables juntas. Estos valores muestran que el contexto visual contiene información sobre la clase: el color del borde, que incluye el fondo y, a veces, parte del fruto, discrimina por encima del azar, y el color global lo hace de forma moderada, en parte porque el oscurecimiento del epicarpio es un síntoma de la BPR. El censo de escenas identificó, además, 22 imágenes del conjunto principal con la mazorca sostenida en la mano sobre un recipiente con otras mazorcas, todas de la clase BPR; ninguna imagen sana del conjunto principal presenta esa composición.

3.2. Dinámica del entrenamiento

La Figura 3 muestra las curvas de la primera repetición. En la primera fase, con el extractor congelado, la pérdida de validación descendió a valores mínimos entre 0.23 y 0.54, según el modelo y el pliegue. La mayoría de las ejecuciones agotó o casi agotó las 20 épocas, salvo ResNet50, cuya pérdida de validación osciló más y que se detuvo en promedio tras 14.6 épocas. Al descongelar la red, la pérdida disminuyó bruscamente y el AUC de validación alcanzó valores entre 0.95 y 1.00. La segunda fase terminó por detención temprana tras 17.6 épocas en promedio en ResNet50 y 30.7 en ConvNeXt-T (Tabla 2). En varios pliegues, la pérdida de validación volvió a subir al final de esta fase, mientras la de entrenamiento seguía bajando, una señal de sobreajuste que la detención temprana contuvo al restaurar los pesos con menor pérdida. Buena parte de la variación entre pliegues se debe a que los conjuntos de validación difirieron en tamaño y en la proporción de clases (Tabla 1).

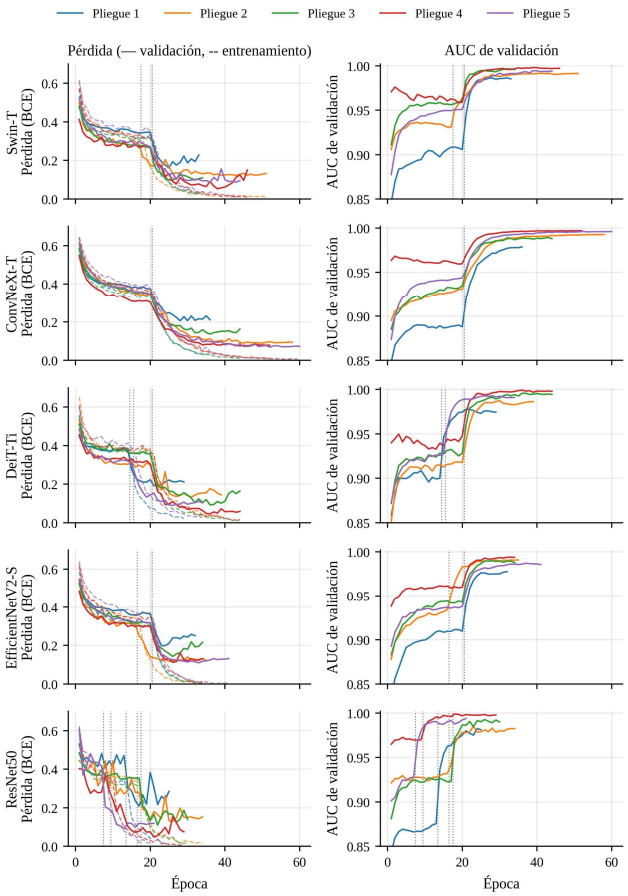


Figura 3. Curvas de pérdida (izquierda) y AUC de validación (derecha) de los cinco pliegues de la repetición 1. Líneas continuas: validación; líneas discontinuas: entrenamiento. La línea vertical punteada marca el inicio de la fase 2 (red completa).

Figure 3. Loss (left) and validation AUC (right) curves for the five folds of repetition 1. Solid lines: validation; dashed lines: training. The dotted vertical line marks the start of stage 2 (full network).

3.3. Desempeño de clasificación

Las cinco arquitecturas alcanzaron un nivel de discriminación interna alto, con diferencias pequeñas en AUC y algo más evidentes en sensibilidad (Tabla 3). El AUC medio por pliegue varió entre 0.989 y 0.993 y la exactitud entre 0.954 y 0.964. Swin-T obtuvo los valores medios más altos (AUC = 0.993 ± 0.004; exactitud = 0.964 ± 0.012;

sensibilidad = 0.959 ± 0.022), seguido de ConvNeXt-T (AUC = 0.992 ± 0.004), que presentó la especificidad media más alta (0.970). DeiT-Ti, EfficientNetV2-S y ResNet50 quedaron por detrás, con sensibilidades de 0.936 a 0.940. Sobre las predicciones fuera de pliegue de la primera repetición, el AUC fue de 0.993 (IC95 %: 0.991 a 0.996) para Swin-T, 0.993 (0.991 a 0.995) para ConvNeXt-T, 0.989 (0.985 a 0.993) para EfficientNetV2-S, 0.988 (0.985 a 0.992) para DeiT-Ti y 0.987 (0.983 a 0.991) para ResNet50, y la exactitud de Swin-T fue de 0.962 (IC95 % bootstrap por grupos: 0.953 a 0.970). El umbral elegido en la validación no mejoró el F1 en comparación con el umbral de 0.5 (Tabla 3).

Las matrices de confusión fuera de pliegue de la primera repetición suman 2211 imágenes en todos los modelos (Figura 4). Swin-T cometió 44 falsos negativos y 39 falsos positivos; ConvNeXt-T, 52 y 31; DeiT-Ti, 69 y 41; EfficientNetV2-S, 70 y 36, y ResNet50, 79 y 31. Swin-T fue así el modelo que dejó sin detectar menos mazorcas enfermas, mientras que ConvNeXt-T y ResNet50 cometieron menos falsos positivos. Las curvas ROC se superponen casi por completo y solo se separan en la región de baja tasa de falsos positivos (Figura 5A).

Tabla 3. Desempeño en los conjuntos de prueba (media ± desviación estándar, en 15 ejecuciones: 3 repeticiones × 5 pliegues).
Table 3. Performance on the test sets (mean ± standard deviation of 15 runs: 3 repetitions × 5 folds).

Métrica	Swin-T	ConvNeXt-T	DeiT-Ti	EfficientNetV2-S	ResNet50
AUC	0.993 ± 0.004	0.992 ± 0.004	0.989 ± 0.005	0.990 ± 0.004	0.990 ± 0.005
AP	0.992 ± 0.005	0.992 ± 0.004	0.989 ± 0.006	0.989 ± 0.006	0.989 ± 0.007
Exactitud	0.964 ± 0.012	0.960 ± 0.012	0.954 ± 0.013	0.956 ± 0.013	0.957 ± 0.012
Precisión	0.959 ± 0.017	0.964 ± 0.015	0.959 ± 0.017	0.960 ± 0.016	0.963 ± 0.026
Sensibilidad	0.959 ± 0.022	0.946 ± 0.021	0.936 ± 0.026	0.940 ± 0.028	0.939 ± 0.026
Especificidad	0.963 ± 0.022	0.970 ± 0.014	0.965 ± 0.017	0.964 ± 0.025	0.968 ± 0.021
F1	0.959 ± 0.015	0.955 ± 0.013	0.947 ± 0.018	0.950 ± 0.018	0.951 ± 0.017
F1 macro	0.961 ± 0.014	0.958 ± 0.012	0.951 ± 0.015	0.953 ± 0.015	0.954 ± 0.014
F1 (umbral de validación)	0.955 ± 0.018	0.949 ± 0.023	0.940 ± 0.016	0.944 ± 0.022	0.949 ± 0.018
Brier	0.029 ± 0.009	0.032 ± 0.010	0.037 ± 0.010	0.035 ± 0.010	0.034 ± 0.009
ECE	0.027 ± 0.008	0.027 ± 0.011	0.031 ± 0.011	0.030 ± 0.008	0.030 ± 0.009

Nota. Métricas de la clase BPR con umbral de 0.5, salvo F1 (umbral de validación). AP: precisión promedio. La DE resume la variabilidad entre 15 ejecuciones dependientes y no equivale a un intervalo de confianza.
Note. BPR class metrics with a 0.5 threshold, except for F1 (validation threshold). AP: average precision. The SD summarizes the variation across 15 dependent runs and does not represent a confidence interval.

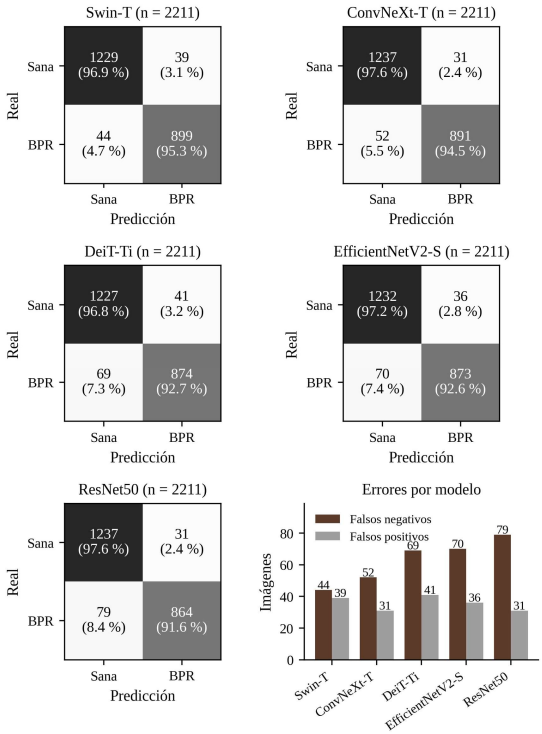


Figura 4. Matrices de confusión fuera de pliegue (repetición 1, umbral 0.5). Cada matriz suma las 2211 imágenes, porque cada imagen se evalúa una vez en el pliegue en que fue de prueba. El último panel resume los falsos negativos y falsos positivos de cada modelo.
Figure 4. Out-of-fold confusion matrices (repetition 1, threshold 0.5). Each matrix sums the 2,211 images, because each image is evaluated once in the fold in which it was held out. The last panel summarizes false negatives and false positives for each model.

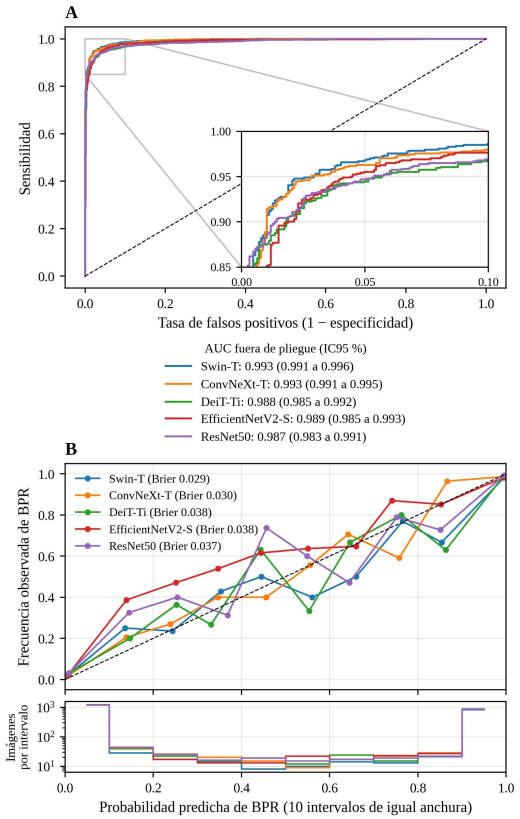


Figura 5. Curvas ROC fuera de pliegue con AUC e IC95 % de DeLong (A; el recuadro amplía la región de baja tasa de falsos positivos) y diagrama de fiabilidad con 10 intervalos de igual anchura y el puntaje de Brier de cada modelo (B), repetición 1. El panel inferior de B indica el número de imágenes por intervalo (escala logarítmica).

Figure 5. Out-of-fold ROC curves with AUC and DeLong 95% CI (A; the inset enlarges the low false positive rate region) and reliability diagram with 10 equal-width bins and each model's Brier score (B), repetition 1. The lower panel of B shows the number of images per bin (logarithmic scale).

En la distribución evaluada, la calibración fue favorable: el puntaje de Brier medio varió entre 0.029 y 0.037 y el ECE entre 0.027 y 0.031 (Tabla 3). La mayoría de las predicciones se concentraron en los intervalos extremos; en los intermedios, con pocas imágenes en cada uno, la frecuencia observada se desvió más de la diagonal (Figura 5B).

En el estrato A, cuyas clases están equilibradas, el AUC fue de 0.986 a 0.991, con especificidades de 0.961 a 0.965 y sensibilidades de 0.930 a 0.957. En el estrato B, los modelos clasificaron correctamente entre 98.4 y 99.3 % de las 368 mazorcas sanas; con solo 14 mazorcas con BPR, la sensibilidad en ese estrato (0.95 a 1.00) es muy imprecisa. Esta evaluación estratificada no equivale a una validación entre fuentes independientes, porque los modelos se entrenaron con ambos estratos.

3.4. Comparaciones estadísticas

La Tabla 4 resume las comparaciones por pares en las tres repeticiones. Las diferencias medias de AUC entre modelos no superaron 0.006, y la mayor diferencia en una repetición individual fue de 0.007. Solo dos ventajas fueron significativas, según la prueba de DeLong con corrección de Holm, en dos de las tres repeticiones: las de Swin-T y ConvNeXt-T sobre ResNet50. Las de Swin-T y ConvNeXt-T sobre DeiT-Ti lo fueron en una sola repetición, y la diferencia entre Swin-T y ConvNeXt-T no lo fue en ninguna. En términos de exactitud, la prueba de McNemar solo detectó diferencias en una repetición: entre Swin-T y ConvNeXt-T frente a DeiT-Ti. Estos resultados describen diferencias pequeñas cuya significación varía entre particiones; no demuestran que los modelos sean equivalentes, porque no se realizó una prueba de equivalencia, ni que alguno sea superior en general.

3.5. Costo computacional

Las diferencias de costo fueron mucho mayores que las de desempeño (Tabla 5). DeiT-Ti tiene 5.5 millones de parámetros, cinco veces menos que Swin-T y ConvNeXt-T, requiere 2.5 GFLOPs por imagen, procesa unas 15 000 imágenes por segundo en GPU con lotes de 64 y tarda 8.1 ms por imagen en CPU. Swin-T necesitó 3.7 ms por imagen en GPU con precisión simple y 41.8 ms en CPU, esta última con una variación considerable entre las mediciones. EfficientNetV2-S fue el más lento en GPU con un lote de 1 (6.5 ms), pese a tener menos FLOPs que Swin-T, ConvNeXt-T y ResNet50. ConvNeXt-T y ResNet50 ocuparon una posición intermedia, con latencias en GPU de 2.2 y 3.5 ms, respectivamente, en precisión simple.

Tabla 4. Comparaciones por pares de las predicciones fuera de pliegue de las tres repeticiones: prueba de DeLong para el AUC y prueba de McNemar para la exactitud (umbral de 0.5), con corrección de Holm.
Table 4. Pairwise comparisons using the out-of-fold predictions of the three repetitions: DeLong test for AUC and McNemar test for accuracy (threshold 0.5), with Holm correction.

Comparación (A frente a B)	ΔAUC medio	p Holm DeLong (mín. a máx.)	Rep. con p < 0.05	ΔExactitud media	Rep. con p < 0.05 (McNemar)
Swin-T frente a ConvNeXt-T	0.0013	0.238 a 1.000	0 de 3	0.0024	0 de 3
Swin-T frente a DeiT-Ti	0.0047	0.032 a 0.238	1 de 3	0.0100	1 de 3
Swin-T frente a EfficientNetV2-S	0.0033	0.073 a 0.956	0 de 3	0.0078	0 de 3
Swin-T frente a ResNet50	0.0059	0.011 a 0.202	2 de 3	0.0069	0 de 3
ConvNeXt-T frente a DeiT-Ti	0.0035	0.032 a 1.000	1 de 3	0.0075	1 de 3
ConvNeXt-T frente a EfficientNetV2-S	0.0021	0.078 a 1.000	0 de 3	0.0054	0 de 3
ConvNeXt-T frente a ResNet50	0.0046	0.003 a 0.238	2 de 3	0.0045	0 de 3
DeiT-Ti frente a EfficientNetV2-S	−0.0014	1.000 a 1.000	0 de 3	−0.0021	0 de 3
DeiT-Ti frente a ResNet50	0.0011	0.630 a 1.000	0 de 3	−0.0030	0 de 3
EfficientNetV2-S frente a ResNet50	0.0025	0.129 a 1.000	0 de 3	−0.0009	0 de 3

Nota. AUC fuera de pliegue (IC95 % de DeLong), repetición 1: Swin-T 0.993 (0.991 a 0.996); ConvNeXt-T 0.993 (0.991 a 0.995); DeiT-Ti 0.988 (0.985 a 0.992); EfficientNetV2-S 0.989 (0.985 a 0.993); ResNet50 0.987 (0.983 a 0.991). Los contrastes son aproximados porque las predicciones fuera de pliegue provienen de modelos entrenados con conjuntos de datos solapados. Δ positivo: A supera a B.
Note. Out-of-fold AUC (DeLong 95% CI), repetition 1: Swin-T 0.993 (0.991 to 0.996); ConvNeXt-T 0.993 (0.991 to 0.995); DeiT-Ti 0.988 (0.985 to 0.992); EfficientNetV2-S 0.989 (0.985 to 0.993); ResNet50 0.987 (0.983 to 0.991). Comparisons are approximate because out-of-fold predictions originate from models trained on overlapping datasets. Positive Δ: A outperforms B.

Tabla 5. Complejidad computacional de los modelos para una imagen de 224 × 224 píxeles.
Table 5. Computational complexity of the models for a 224 × 224 pixel image.

Modelo	Parámetros (M)	GFLOPs	GPU FP32 (ms/img)	GPU BF16 (ms/img)	Rendimiento BF16 (img/s)	Memoria GPU (MB)	CPU 4 hilos (ms/img)
Swin-T	27.52	8.98	3.67 ± 0.18	4.14 ± 0.19	2834 ± 8	133	41.8 ± 12.6
ConvNeXt-T	27.82	8.91	2.18 ± 0.00	2.08 ± 0.07	3970 ± 6	131	25.9 ± 4.7
DeiT-Ti	5.52	2.51	1.70 ± 0.00	1.84 ± 0.01	15074 ± 1	32	8.1 ± 0.7
EfficientNetV2-S	20.18	5.70	6.49 ± 0.02	6.37 ± 0.01	2970 ± 13	102	34.5 ± 2.2
ResNet50	23.51	8.17	3.51 ± 0.00	2.30 ± 0.02	4753 ± 7	126	23.5 ± 2.4

Nota. Media ± DE de tres mediciones, con cada modelo en un proceso independiente. Latencias con lote de 1 en una GPU NVIDIA A100 (float32 sin TF32 y bfloat16) y en una CPU Intel Xeon Gold 5420+ con 4 hilos; rendimiento con lotes de 64; memoria: pico asignado por PyTorch con lote de 1. Se usaron tensores sintéticos, sin carga, decodificación ni preprocesamiento de imágenes. FLOPs contados con una multiplicación-suma como dos operaciones.
Note: Mean ± SD of three measurements, with each model run in an independent process. Latencies measured with a batch size of 1 on an NVIDIA A100 GPU (float32 without TF32, and bfloat16) and on an Intel Xeon Gold 5420+ CPU with four threads; throughput measured with a batch size of 64; memory: peak allocated by PyTorch with a batch size of 1. Synthetic tensors were used; no image loading, decoding, or preprocessing. FLOPs were calculated counting a multiply-accumulate operation as two operations.

3.6. Mapas de activación

En los casos inspeccionados con BPR, los mapas Grad-CAM de las cinco redes se superpusieron al fruto y, en la mayoría de los casos, al tejido necrótico visible (Figura 6). En las mazorcas sanas, Swin-T, ConvNeXt-T, EfficientNetV2-S y ResNet50 activaron principalmente el fruto, mientras que los mapas de DeiT-Ti fueron dispersos y, en dos de las tres imágenes sanas, se extendieron al fondo. Entre los errores de Swin-T (Figura 7) hay dos mazorcas sanas clasificadas como BPR: una con manchas oscuras junto al pedúnculo, que también confundieron EfficientNetV2-S y ResNet50, y otra envuelta en una bolsa plástica con pliegues y reflejos, que además confundió ConvNeXt-T. Los dos falsos negativos fueron una mazorca roja sobre el tronco, que ninguno de los cinco modelos detectó, y un racimo de mazorcas en el que EfficientNetV2-S y ResNet50 sí identificaron la mazorca enferma.

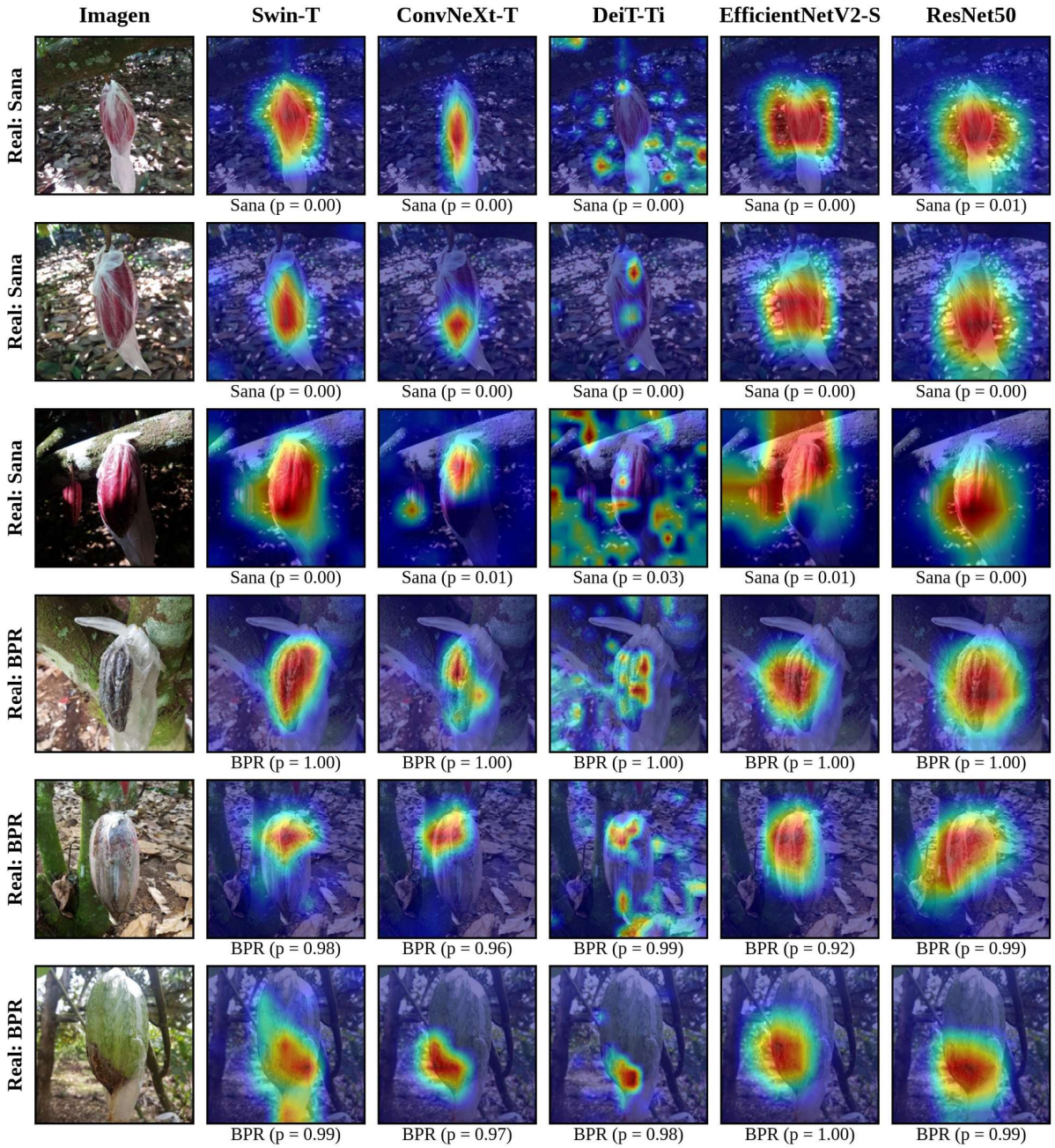


Figura 6. Mapas Grad-CAM de los cinco modelos (repetición 1, pliegue 1) para tres mazorcas sanas y tres con BPR elegidas al azar del conjunto de prueba. Cada mapa resalta la evidencia a favor de la clase predicha (para la clase sana se usa el logit con signo negativo). Bajo cada mapa se indica la clase predicha y la probabilidad de BPR. Análisis cualitativo de casos.

Figure 6. Grad-CAM maps of the five models (repetition 1, fold 1) for three healthy and three BPR pods randomly drawn from the test set. Each map highlights the evidence for the predicted class (the negative logit is used for the healthy class). The predicted class and BPR probability are shown below each map. Qualitative case analysis.

3.7. Evaluación con imágenes sanas no utilizadas

El control de reproducción confirmó que los modelos se aplicaron bajo las mismas condiciones que en la validación cruzada: al repetir la predicción en cada pliegue de prueba, la diferencia con las probabilidades originales fue de 7×10^{-5} en promedio y de 0.0068 como máximo. De las 2076 sanas no

utilizadas, 118 estaban vinculadas por secuencia de captura al conjunto principal y se excluyeron; por lo tanto, se evaluaron 1958. En ellas, la especificidad fue de 0.952 a 0.965, entre 0.0 y 1.4 puntos porcentuales por debajo de la obtenida en la validación cruzada (Tabla 6). La diferencia solo excluyó el cero en dos repeticiones para ConvNeXt-T y en una para EfficientNetV2-S. Al excluir también las imágenes con vínculos dudosos, la especificidad se mantuvo prácticamente igual (de 0.952 a 0.965). El AUC con nuevos negativos fue de 0.987 a 0.993.

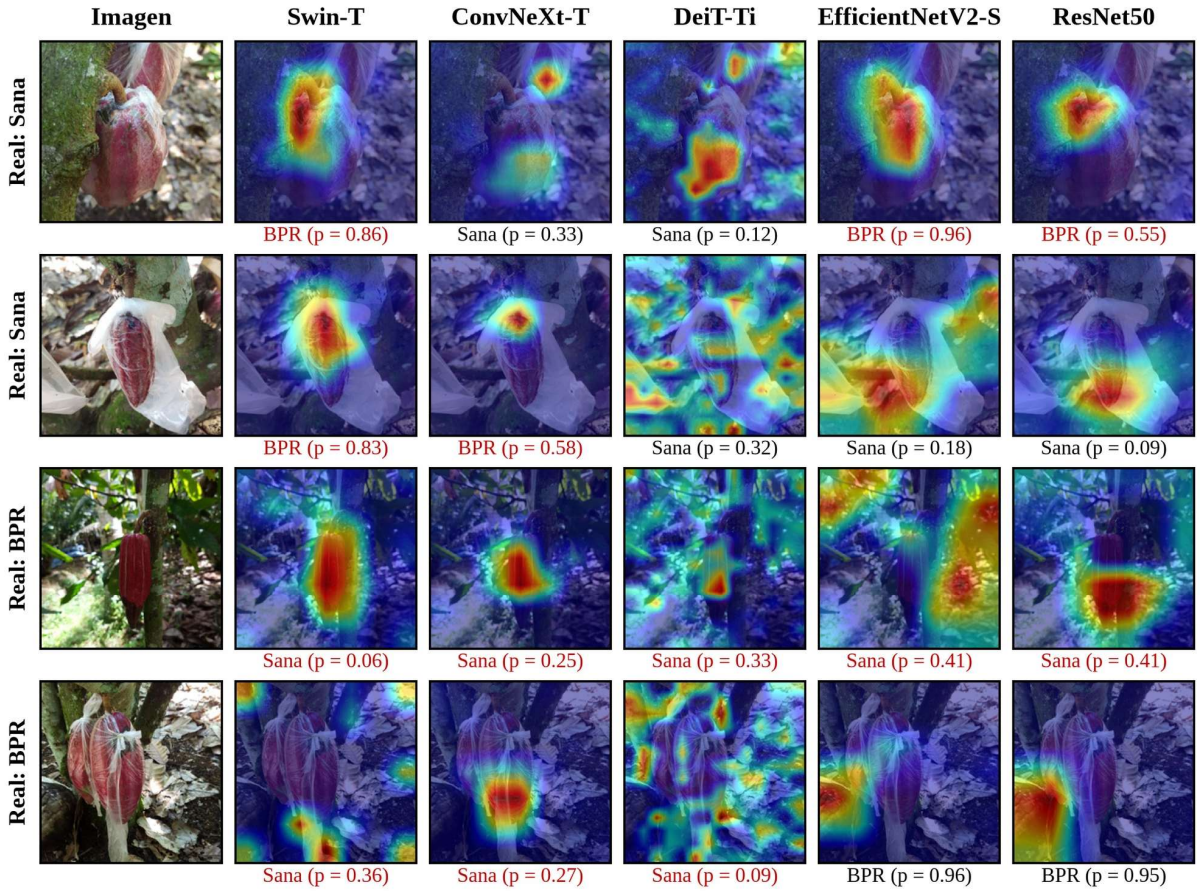


Figura 7. Mapas Grad-CAM de las imágenes mal clasificadas por Swin-T (repetición 1, pliegue 1: dos falsos positivos y dos falsos negativos) y las respuestas de los demás modelos para esas mismas imágenes. Las predicciones erróneas se indican en rojo.
Figure 7. Grad-CAM maps of the images misclassified by Swin-T (repetition 1, fold 1: two false positives and two false negatives) and the response of the other models to the same images. Incorrect predictions are shown in red.

Tabla 6. Especificidad y AUC en las imágenes sanas del repositorio no utilizadas en el conjunto principal, frente a la validación cruzada (media ± desviación estándar de los 15 modelos entrenados de cada arquitectura).
Table 6. Specificity and AUC on the healthy repository images not used in the main dataset, compared with cross-validation (mean ± standard deviation of the 15 trained models of each architecture).

Métrica	Swin-T	ConvNeXt-T	DeiT-Ti	EfficientNetV2-S	ResNet50
Especificidad, validación cruzada	0.963 ± 0.022	0.970 ± 0.014	0.965 ± 0.017	0.964 ± 0.025	0.968 ± 0.021
Especificidad, sanas no utilizadas (n = 1958)	0.961 ± 0.018	0.956 ± 0.017	0.965 ± 0.012	0.952 ± 0.013	0.960 ± 0.017
Escena mano + recipiente (n = 89)	0.727 ± 0.210	0.519 ± 0.239	0.849 ± 0.104	0.659 ± 0.134	0.665 ± 0.182
Resto de escenas (n = 1869)	0.972 ± 0.011	0.977 ± 0.009	0.971 ± 0.009	0.966 ± 0.009	0.974 ± 0.011
Sin vínculos dudosos (n = 1442)	0.961 ± 0.015	0.965 ± 0.013	0.964 ± 0.012	0.952 ± 0.013	0.964 ± 0.015
Falsos positivos en sanas no utilizadas (media)	77.0	86.5	67.7	94.1	77.9
Repeticiones con diferencia de especificidad[a]	0 de 3	2 de 3	0 de 3	1 de 3	0 de 3
AUC, validación cruzada	0.993 ± 0.004	0.992 ± 0.004	0.989 ± 0.005	0.990 ± 0.004	0.990 ± 0.005
AUC con negativos nuevos[b]	0.993 ± 0.004	0.990 ± 0.005	0.988 ± 0.006	0.987 ± 0.006	0.988 ± 0.006

Nota. Umbral 0.5. De las 2076 sanas no utilizadas, se excluyeron 118 vinculadas por secuencia de captura al conjunto principal. Escena mano + recipiente: mazorca sostenida en la mano sobre un recipiente con otras mazorcas, identificada por similitud DINOv2 (≥ 0.80) con el centroide de las 22 imágenes de esa escena en el conjunto principal, todas de la clase BPR. Sin vínculos dudosos: se excluyen también las imágenes con similitud intermedia respecto al conjunto principal. [a]Número de repeticiones en que el IC95 % bootstrap por grupos de la diferencia entre la especificidad en las sanas no utilizadas y la especificidad fuera de pliegue no incluye el cero. [b]Positivos: imágenes con BPR del pliegue de prueba de cada modelo, ya usadas en la evaluación principal; negativos: sanas no utilizadas.
Note. Threshold 0.5. Of the 2076 unused healthy images, 118 linked by capture sequence to the main set were excluded. Hand + container scene: pod held in hand over a container with other pods, identified by DINOv2 similarity (≥ 0.80) with the centroid of the 22 images of that scene in the main set, all of the BPR class. No dubious links: images with intermediate similarity to the main set are also excluded. [a]Number of repetitions in which the group bootstrap 95% CI of the difference between the specificity on the unused healthy images and the out-of-fold specificity does not include zero. [b]Positives: BPR images of the test fold of each model, already used in the main evaluation; Negatives: unused healthy images.

La especificidad dependió de la escena de captura. Las 89 sanas no utilizadas con la mazorca en la mano sobre un recipiente, la composición que en el entrenamiento solo aparecía en mazorcas con BPR, obtuvieron especificidades medias de 0.519 (ConvNeXt-T), 0.659 (EfficientNetV2-S), 0.665 (ResNet50), 0.727 (Swin-T) y 0.849 (DeiT-Ti), con una gran variación entre los modelos entrenados de cada arquitectura. En las 1869 imágenes restantes, la especificidad fue de 0.966 a 0.977, similar o superior a la obtenida en la validación cruzada. El orden de robustez en esta escena no coincidió con el orden del AUC: DeiT-Ti, con el AUC más bajo en la validación cruzada, fue el menos afectado, y ConvNeXt-T, el segundo en AUC, el más afectado.

4. DISCUSIÓN

Las cinco arquitecturas discriminaron muy bien entre mazorcas sanas y con BPR. Las diferencias de AUC fueron de milésimas, y solo las de Swin-T y ConvNeXt-T frente a ResNet50 resultaron significativas en dos de las tres repeticiones. Sería un error leer esto como equivalencia, ya que no se fijó un margen que la definiera; lo que sí puede afirmarse es que, con este protocolo, cambiar de arquitectura apenas modifica la discriminación. Navarro-Cabrera et al. (2026) llegaron a una conclusión similar con ResNet50, EfficientNet-B0 y ViT-B/16 en imágenes de cacao, y el resultado concuerda con lo observado en otros problemas de transferencia con pocos miles de imágenes, en los que la receta de entrenamiento influye tanto como el diseño de la red (KORNBLITH et al., 2019; LIU et al., 2022).

Buena parte del trabajo se dedicó a garantizar que el entrenamiento y la prueba no compartieran fotografías de la misma mazorca. El repositorio está formado en gran medida por series de tomas: el 79 % de las imágenes tiene al menos una compañera de grupo, y muchas de esas tomas difieren solo en el ángulo o el encuadre. Este parentesco no se detecta ni al revisar únicamente los pares más parecidos ni con umbrales estrictos de similitud, y si se ignora, el modelo puede acertar porque reconoce la mazorca y no la enfermedad (KAPOOR; NARAYANAN, 2023). Por esa razón todas las estimaciones se obtuvieron con particiones por secuencia de captura, aun a costa de pliegues de tamaño desigual. Es probable que otros conjuntos de imágenes de campo, compuestos también por series de fotografías, tengan una estructura similar (Barbedo, 2022), y que las exactitudes obtenidas en ellos mediante particiones por imagen deban interpretarse con cuidado.

Dentro de un margen tan estrecho, lo que distinguió a los modelos fue el tipo de error. Swin-T dejó sin detectar menos mazorcas enfermas, y para el manejo del cultivo esa es la propiedad más valiosa, porque una mazorca con BPR que permanece en el árbol sigue produciendo inóculo (GUEST, 2007). Su atención por ventanas desplazadas integra información local y contextual en varias escalas (Liu et al., 2021), aunque con estos datos no es posible atribuirle la ventaja. ConvNeXt-T cometió menos falsos positivos y alcanzó un AUC casi idéntico con menor latencia en GPU, lo que apoya la idea de que una red convolucional modernizada puede competir con los transformers jerárquicos (LIU et al., 2022).

DeiT-Ti, con 5.5 millones de parámetros, fue el más ligero y el más rápido en el servidor. Es un buen candidato para probar en teléfonos o dispositivos de borde, a cambio de una sensibilidad algo menor, pero su idoneidad dependerá de mediciones en el equipo de destino, que no se realizaron aquí.

Los resultados publicados con EfficientNet en el cacao son muy dispares. Pradana et al. (2024) obtuvieron 97.3 % de exactitud con EfficientNet-B4 tras aumentar los datos, y Maylianti et al. (2025), 96 % con EfficientNet-B0. EfficientNetV2-S alcanzó aquí 95.6 % de exactitud y un AUC de 0.990, cifras coherentes con los resultados de esos trabajos. Hay dos detalles de implementación que afectan especialmente a esta familia y, por eso, se declaran. En timm, la capa de salida de EfficientNet se inicializa con valores uniformes en [−1, 1] cuando tiene una sola neurona, lo que produce logits iniciales muy grandes; para evitarlo se usó la misma inicialización normal truncada en los cinco modelos. Además, las capas de normalización por lotes se mantuvieron en modo de inferencia al descongelar la red, porque recalcular sus estadísticas con lotes de 32 imágenes altera de golpe las características que recibe la cabecera ya entrenada (IOFFE; SZEGEDY, 2015). Como no se realizaron ablaciones, no es posible cuantificar cuánto pesó cada decisión en el resultado final.

Los valores obtenidos se sitúan dentro del rango reportado para el cacao (Tabla 7). Maylianti et al. (2025) obtuvieron 96 % de exactitud con ResNet50 y EfficientNet-B0 en el mismo repositorio; Rola et al. (2024) informaron 99 % con una CNN entrenada con 2000 imágenes propias; y Techie-Menson et al. (2026), 98.95 % en un conjunto de Ghana con tres clases. Con fotografías de plantaciones de San Martín, Perú, Villalobos-Culqui et al. (2025) obtuvieron cerca del 79 % con YOLOv8 e InceptionV3; y Yanac-Montesino, Rosales-Huamani; Castillo-Sequera (2021), del 83 % con ResNet18. Estas cifras no son directamente comparables. Cambian el número de clases, el origen de las imágenes y la forma de evaluar, que en varios casos se reduce a una sola partición por imagen, sin intervalos de confianza ni control de fotografías repetidas. La distancia entre los estudios con imágenes de repositorio y los que toman sus propias fotografías en campo es un motivo para leer con cautela las cifras superiores al 95 %. El caso de Gado; Primandari (2025) es elocuente: su modelo alcanzó 94.9 % de exactitud en prueba, pero en una aplicación móvil utilizada en campo, la precisión de la clase enferma bajó a 57.1 %. Ahmad et al. (2023) describen la misma caída al pasar de condiciones controladas a condiciones de campo en otros cultivos.

Los controles de atajos y la evaluación por escena indican que no todo lo aprendido corresponde a la lesión. Rasgos tan simples como el color del borde o el color global ya predicen la clase por encima del azar, y que las redes los superen con holgura no descarta que aprovechen atajos más elaborados (GEIRHOS et al., 2020). La escena de la mazorca en la mano sobre un recipiente es un ejemplo claro. En el conjunto principal solo aparecía en imágenes con BPR, y, ante mazorcas sanas fotografiadas de ese modo, la especificidad bajó a valores entre 0.52 y 0.85, mientras que en el resto de escenas se mantuvo cerca de 0.97. Nada en el entrenamiento permitía a los modelos separar esa escena de la enfermedad. Zech et al. (2018) describieron algo similar en radiografías de tórax, donde un modelo para detectar neumonía perdió rendimiento en hospitales externos porque se basaba en características propias del lugar de adquisición. En plantas,

Mohanty et al. (2016) observaron que la exactitud de su modelo caía a poco más del 31 % con imágenes tomadas en condiciones distintas de las de entrenamiento, y Barbedo (2022) relaciona estos fallos con conjuntos de datos que no cubren la variabilidad real de la captura. El caso del recipiente añade un matiz: no basta con que un contexto esté representado, tiene que estarlo en ambas clases.

Tabla 7. Estudios recientes sobre la clasificación o detección de enfermedades de la mazorca de cacao mediante aprendizaje profundo.
Table 7. Recent studies on the classification or detection of cacao pod diseases with deep learning.

Estudio	Datos (imágenes; clases)	Modelo	Resultado reportado
Yanac Montesino et al. (2021)	Imágenes de mazorcas; sana y *P. palmivora*	ResNet18	Exactitud ≈ 83 %
Pradana et al. (2024)	Repositorio Kaggle; BPR y barrenador	EfficientNet-B4	Exactitud 97.3 %
Rola et al. (2024)	2000 propias; 2 clases	CNN	Exactitud 99 %
Maylianti et al. (2025)	4390 (repositorio de Pagaduan); 3 clases	EfficientNet-B0; ResNet50	Exactitud 96 %
Villalobos-Culqui et al. (2025)	1982 propias, San Martín, Perú	YOLOv8; InceptionV3	Exactitud ≈ 79 %
Sudjud et al. (2025)	4291 propias y de Kaggle; 2 clases	YOLOv9	mAP 0.99
Gado y Primandari (2025)	2000; 4 clases	NASNet-Mobile	Exactitud 94.9 % (prueba); precisión 57.1 % en campo
Navarro-Cabrera et al. (2026)	4390 (repositorio de Pagaduan); 3 clases; VC 5 pliegues	ResNet50; EfficientNet-B0; ViT-B/16	F1 macro > 0.96, sin diferencias significativas
Techie-Menson et al. (2026)	1704 propias, Ghana; 3 clases	CNN con LGF-CBAM	Exactitud 98.95 %
Este estudio	2211 (repositorio de Pagaduan); 2 clases; VC 3 × 5 por secuencia de captura	Swin-T; ConvNeXt-T; DeiT-Ti; EfficientNetV2-S; ResNet50	AUC 0.989 a 0.993; exactitud 0.954 a 0.964

Nota. VC: validación cruzada. Los resultados no son directamente comparables porque difieren en los datos, las clases y los esquemas de evaluación.
Note. CV: cross-validation. The results are not directly comparable because the data, classes, and evaluation schemes differ.

Los mapas Grad-CAM no permitieron anticipar este problema. En los casos inspeccionados, la activación recaía casi siempre en el fruto y, aun así, los modelos dependían de la escena en un subconjunto que unos pocos ejemplos no podían revelar. Tampoco cabía esperarlo de ellos, puesto que una explicación visual no demuestra que la región resaltada cause la decisión (ADEBAYO et al., 2018). Llama la atención, además, que la robustez frente a la escena no siguiera el orden del AUC: DeiT-Ti fue el menos afectado y ConvNeXt-T, el que más, de modo que la elección de un modelo para el campo debería incluir pruebas de este tipo. Fuera de esa escena y una vez excluidas las sanas vinculadas por secuencia al conjunto principal, la especificidad en imágenes nuevas del repositorio coincidió con la estimada mediante validación cruzada. Recht et al. (2019) observaron caídas generalizadas en la exactitud al construir nuevos conjuntos de prueba para ImageNet; aquí, el deterioro se concentró en un contexto mal representado, lo que señala con bastante precisión dónde habría que ampliar los datos.

En la distribución evaluada, las probabilidades estuvieron razonablemente bien calibradas, con puntajes de Brier de 0.029 a 0.037 y de ECE de 0.027 a 0.031. Eso no garantiza que lo estén en campo. La prevalencia de BPR en el conjunto analizado (42.7 %) quedó fijada al reducir la clase sana, mientras que en las dos clases del repositorio completo ronda el 22 %, y en una plantación puede ser muy distinta; además, las redes profundas tienden a volverse sobreconfiadas cuando cambian las condiciones (GUO et al., 2017). Antes de usar las probabilidades para decidir, habría que recalibrarlas con datos de la población de destino. También conviene separar dos usos posibles: ordenar la inspección empezando por las mazorcas con mayor probabilidad de BPR, o revisar a mano las de probabilidad intermedia, en las que el modelo es menos fiable. Si un falso negativo se considera más costoso que un falso positivo, el umbral debería fijarse en la validación interna a partir de una sensibilidad objetivo o de un costo esperado, y no maximizando el F1, que en este estudio no mejoró respecto del umbral 0.5.

4.1. Limitaciones

Los resultados deben interpretarse con ciertas limitaciones. Primero, todas las imágenes proceden de un único repositorio público, sin información sobre su origen geográfico, el cultivar, la fecha de captura ni la cámara. No hubo validación externa: la evaluación con las sanas no utilizadas proviene del mismo repositorio y solo permite estimar la especificidad; los estratos A y B se definieron por resolución y no por fuente confirmada; y la escena del recipiente se identificó a partir de una composición observada durante la auditoría, no de una hipótesis preespecificada. Segundo, las etiquetas no fueron verificadas por un fitopatólogo; algunas sanas clasificadas como BPR presentaban manchas oscuras que podrían ser lesiones iniciales, y el repositorio no identifica la especie de *Phytophthora*. Dado que *P. palmivora*, *P. megakarya* y *P. capsici* producen síntomas similares, pero presentan distribuciones y epidemiologías distintas (Guest, 2007; Marelli et al., 2019), los modelos no deben usarse para inferir la especie. Tercero, el enfoque es binario: no distingue grados de severidad ni otras enfermedades de la mazorca, como la moniliasis, frecuente en el Perú, ni el daño causado por el barrenador.

Cuarto, la agrupación por secuencia de captura reduce la fuga, pero no la elimina: según la proporción observada en la revisión parcial, unos 44 pares de la misma mazorca o escena podrían no haberse agrupado, y los umbrales también podrían haber unido mazorcas distintas dentro de una misma escena. Además, los grupos de gran tamaño produjeron pliegues de tamaño y composición desiguales. Quinto, la validación cruzada repetida utiliza modelos entrenados con conjuntos de datos solapados, de modo que los intervalos y contrastes resultantes son aproximados y no capturan toda la variabilidad del reentrenamiento (BATES et al., 2024). Sexto, el subconjunto sano se formó con tramos de numeración cuyo sorteo no quedó registrado, lo que modificó la prevalencia en comparación con la del repositorio completo. Séptimo, no se realizaron ablaciones de las decisiones de implementación. Por último, los tiempos de inferencia se midieron en un servidor con tensores sintéticos, sin la carga

ni el preprocesamiento de imágenes, y no en teléfonos ni en vehículos aéreos no tripulados. Los trabajos futuros deberán validar los modelos con imágenes de otras plantaciones, incluidas las peruanas, con sanas y enfermas, en los mismos contextos de captura, etiquetadas por especialistas y con varios niveles de severidad, y medir el desempeño directamente en el dispositivo de uso.

5. CONCLUSIONES

Entrenadas con un protocolo común y evaluadas mediante validación cruzada repetida por grupos, Swin-T, ConvNeXt-T, DeiT-Ti, EfficientNetV2-S y ResNet50 alcanzaron una discriminación interna alta, con AUC medios entre 0.989 y 0.993 y diferencias pequeñas entre sí. Swin-T tuvo las mejores métricas medias y la mayor sensibilidad; ConvNeXt-T, la mayor especificidad; y DeiT-Ti, el menor costo en el hardware evaluado. Las diferencias de AUC solo fueron significativas en algunas comparaciones y repeticiones, por lo que no permiten afirmar la superioridad de una arquitectura ni su equivalencia.

La mayoría de las imágenes del repositorio forma parte de series de tomas de una misma mazorca o escena, de modo que la independencia entre entrenamiento y prueba debe controlarse antes de comparar modelos. En imágenes sanas nuevas del mismo repositorio, la especificidad coincidió con la de la validación cruzada, salvo en una escena que, en el entrenamiento, solo aparecía en mazorcas enfermas, donde bajo a valores de 0.52 a 0.85. Los modelos aprendieron, por tanto, asociaciones contextuales además de los rasgos de la lesión.

Estos resultados no sustituyen una validación externa. Antes de llevar los modelos al campo habrá que evaluarlos con imágenes de otras plantaciones, con mazorcas sanas y enfermas fotografiadas en los mismos contextos, etiquetadas por especialistas y medidas en el dispositivo de uso. Sería deseable, además, que los estudios sobre enfermedades del cacao describieran el protocolo completo, auditaran la identidad de sus imágenes y acompañaran las métricas con su incertidumbre.

6. REFERENCIAS

ABADE, A.; FERREIRA, P. A.; VIDAL, F. de B. Plant diseases recognition on images using convolutional neural networks: a systematic review. **Computers and Electronics in Agriculture**, v. 185, e106125, 2021. https://doi.org/10.1016/j.compag.2021.106125

ADEBAYO, J.; GILMER, J.; MUELLY, M.; GOODFELLOW, I.; HARDT, M.; KIM, B. Sanity checks for saliency maps. In: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 32, 2018, Montreal. **Advances in Neural Information Processing Systems**, v. 31, p. 9505-9515, 2018. Disponible en: https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html

AHMAD, A.; GAMAL, A. E.; SARASWAT, D. Toward generalization of deep learning-based plant disease identification under controlled and field conditions. **IEEE Access**, v. 11, p. 9042-9057, 2023. https://doi.org/10.1109/ACCESS.2023.3240100

BARBEDO, J. G. A. Deep learning applied to plant pathology: the problem of data representativeness. **Tropical Plant Pathology**, v. 47, n. 1, p. 85-94, 2022. https://doi.org/10.1007/s40858-021-00459-9

BARBEDO, J. G. A. Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. **Computers and Electronics in Agriculture**, v. 153, p. 46-53, 2018. https://doi.org/10.1016/j.compag.2018.08.013

BATES, S.; HASTIE, T.; TIBSHIRANI, R. Cross-validation: what does it estimate and how well does it do it? **Journal of the American Statistical Association**, v. 119, n. 546, p. 1434-1445, 2024. https://doi.org/10.1080/01621459.2023.2197686

BEAKES, G. W.; GLOCKLING, S. L.; SEKIMOTO, S. The evolutionary phylogeny of the oomycete "fungi". **Protoplasma**, v. 249, n. 1, p. 3-19, 2012. https://doi.org/10.1007/s00709-011-0269-2

BRIER, G. W. Verification of forecasts expressed in terms of probability. **Monthly Weather Review**, v. 78, n. 1, p. 1-3, 1950. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

DELONG, E. R.; DELONG, D. M.; CLARKE-PEARSON, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. **Biometrics**, v. 44, n. 3, p. 837-845, 1988. https://doi.org/10.2307/2531595

DENG, J.; DONG, W.; SOCHER, R.; LI, L.-J.; LI, K.; FEI-FEI, L. ImageNet: a large-scale hierarchical image database. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2009, Miami. **Proceedings** [...]. Piscataway: IEEE, 2009. p. 248-255. https://doi.org/10.1109/CVPR.2009.5206848

DIETTERICH, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. **Neural Computation**, v. 10, n. 7, p. 1895-1923, 1998. https://doi.org/10.1162/089976698300017197

FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861-874, 2006. https://doi.org/10.1016/j.patrec.2005.10.010

FISCHLER, M. A.; BOLLES, R. C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. **Communications of the ACM**, v. 24, n. 6, p. 381-395, 1981. https://doi.org/10.1145/358669.358692

GADO, G. A. S.; PRIMANDARI, P. N. Sistem klasifikasi berbasis Android untuk penyakit buah kakao menggunakan CNN NasNet-Mobile. **Jurnal Teknologi Terpadu**, v. 11, n. 1, p. 27-35, 2025. https://doi.org/10.54914/jtt.v11i1.1821

GEIRHOS, R.; JACOBSEN, J.-H.; MICHAELIS, C.; ZEMEL, R.; BRENDEL, W.; BETHGE, M.; WICHMANN, F. A. Shortcut learning in deep neural networks. **Nature Machine Intelligence**, v. 2, n. 11, p. 665-673, 2020. https://doi.org/10.1038/s42256-020-00257-z

GODMALIN, R. A.; ALIAC, C. J.; FELISCUZO, L. Cacao pod infection level classification using transfer learning. In: IEEE OPEN CONFERENCE OF ELECTRICAL, ELECTRONIC AND INFORMATION SCIENCES (eStream), 2023, Vilnius. **Proceedings** [...]. Piscataway: IEEE, 2023. 6p. https://doi.org/10.1109/eStream59056.2023.10135062

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. Cambridge: MIT Press, 2016. 800p.

GUEST, D. Black pod: diverse pathogens with a global impact on cocoa yield. **Phytopathology**, v. 97, n. 12, p. 1650-1653, 2007. https://doi.org/10.1094/PHYTO-97-12-1650

GUO, C.; PLEISS, G.; SUN, Y.; WEINBERGER, K. Q. On calibration of modern neural networks. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 34, 2017, Sydney. **Proceedings of Machine Learning Research** [...]. v. 70, p. 1321-1330, 2017. Disponible en: https://proceedings.mlr.press/v70/guo17a.html

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2016, Las Vegas. **Proceedings** [...]. Piscataway: IEEE, 2016. p. 770-778. https://doi.org/10.1109/CVPR.2016.90

HOLM, S. A simple sequentially rejective multiple test procedure. **Scandinavian Journal of Statistics**, v. 6, n. 2, p. 65-70, 1979. Disponible en: https://www.jstor.org/stable/4615733

HOWARD, J.; RUDER, S. Universal language model fine-tuning for text classification. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 56, 2018, Melbourne. **Proceedings** [...]. Stroudsburg: ACL, 2018. p. 328-339. https://doi.org/10.18653/v1/P18-1031

IOFFE, S.; SZEGEDY, C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 32, 2015, Lille. **Proceedings of Machine Learning Research**, v. 37, p. 448-456, 2015. Disponible en: https://proceedings.mlr.press/v37/ioffe15.html

KAMOUN, S.; FUZER, O.; JONES, J. D. G.; JUDELSON, H. S.; ALI, G. S.; DALIO, R. J. D.; ROY, S. G.; SCHENA, L.; ZAMBOUNIS, A.; et al. The Top 10 oomycete pathogens in molecular plant pathology. **Molecular Plant Pathology**, v. 16, n. 4, p. 413-434, 2015. https://doi.org/10.1111/mpp.12190

KAPOOR, S.; NARAYANAN, A. Leakage and the reproducibility crisis in machine-learning-based science. **Patterns**, v. 4, n. 9, e100804, 2023. https://doi.org/10.1016/j.patter.2023.100804

KORNBLITH, S.; SHLENS, J.; LE, Q. V. Do better ImageNet models transfer better? In: IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2019, Long Beach. **Proceedings** [...]. Piscataway: IEEE, 2019. p. 2661-2671. https://doi.org/10.1109/CVPR.2019.00277

LIU, Z.; LIN, Y.; CAO, Y.; HU, H.; WEI, Y.; ZHANG, Z.; LIN, S.; GUO, B. Swin Transformer: hierarchical vision transformer using shifted windows. In: IEEE/CVF INTERNATIONAL CONFERENCE ON COMPUTER VISION, 2021, Montreal. **Proceedings** [...]. Piscataway: IEEE, 2021. p. 9992-10002. https://doi.org/10.1109/ICCV48922.2021.00986

LIU, Z.; MAO, H.; WU, C.-Y.; FEICHTENHOFER, C.; DARRELL, T.; XIE, S.; A ConvNet for the 2020s. In: IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2022, New Orleans. **Proceedings** [...]. Piscataway: IEEE, 2022. p. 11966-11976. https://doi.org/10.1109/CVPR52688.2022.01167

LOSHCHILOV, I.; HUTTER, F. Decoupled weight decay regularization. In: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, 7, 2019, New Orleans. **Proceedings** [...]. 2019. Disponible en: https://openreview.net/forum?id=Bkg6RiCqY7

LOWE, D. G. Distinctive image features from scale-invariant keypoints. **International Journal of Computer Vision**, v. 60, n. 2, p. 91-110, 2004. https://doi.org/10.1023/B:VISI.0000029664.99615.94

LUQUE, A.; CARRASCO, A.; MARTÍN, A.; LAS HERAS, A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. **Pattern Recognition**, v. 91, p. 216-231, 2019. https://doi.org/10.1016/j.patcog.2019.02.023

MARELLI, J.-P.; GUEST, D. I.; BAILEY, B. A.; EVANS, H. C.; BROWN, J. K.; JUNAID, M.; BARRETO, R. W.; LISBOA, D. O.; PUIG, A. S. Chocolate under threat from old and new cacao diseases. **Phytopathology**, v. 109, n. 8, p. 1331-1343, 2019. https://doi.org/10.1094/PHYTO-12-18-0477-RVW

MAYLIANTI, N. P.; WIJAYAKUSUMA, I. G. N. L.; WIGUNA, I. P. C. A. Comparison of EfficientNet-B0 and ResNet-50 for detecting diseases in cocoa fruit. **Journal of Applied Informatics and Computing**, v. 9, n. 1, p. 115-120, 2025. https://doi.org/10.30871/jaic.v9i1.8868

MCNEMAR, Q. Note on the sampling error of the difference between correlated proportions or percentages. **Psychometrika**, v. 12, n. 2, p. 153-157, 1947. https://doi.org/10.1007/BF02295996

MOHANTY, S. P.; HUGHES, D. P.; SALATHÉ, M. Using deep learning for image-based plant disease detection. **Frontiers in Plant Science**, v. 7, e1419, 2016. https://doi.org/10.3389/fpls.2016.01419

NAVARRO-CABRERA, J. R.; BERAÚN-BARRANTES, J. G.; CÁRDENAS-GARCÍA, Á.; LOZANO-CARRANZA, C. M. Aplicación de aprendizaje profundo en la detección fitosanitaria del cacao usando visión por computadora. **Revista Científica de Sistemas e Informática**, v. 6, n. 1, e1385, 2026. https://doi.org/10.51252/rcsi.v6i1.1385

OQUAB, M.; DARCET, T.; MOUTAKANNI, T.; VO, H. V.; SZAFRANIEC, M.; KHALIDOV, V.; FERNANDEZ, P.; HAZIZA, D.; MASSA, F.; EL-NOUBY, A.; et al. DINOv2: learning robust visual features without supervision. **Transactions on Machine Learning Research**, p. 1-32, 2024. Disponible en: https://openreview.net/forum?id=a68SUt6zFt

PAGADUAN, Z. Jr. **Cacao Diseases**. Kaggle, 2021. Conjunto de datos. Disponible en: https://www.kaggle.com/datasets/zaldyjr/cacao-diseases

PASZKE, A.; GROSS, S.; MASSA, F.; LERER, A.; BRADBURY, J.; CHANAN, G.; KILLEEN, T.; LIN, Z.; GIMELSHEIN, N.; ANTIGA, L.; et al. PyTorch: an imperative style, high-performance deep learning library. In: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 33, 2019, Vancouver. **Advances in Neural Information Processing Systems**, v. 32, p. 8024-8035, 2019. Disponible en: https://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library

PERRINE-WALKER, F. Phytophthora palmivora-cocoa interaction. **Journal of Fungi**, v. 6, n. 3, 167, 2020. https://doi.org/10.3390/jof6030167

PRADANA, A. G.; SETIADI, D. R. I. M.; MUSLIKH, A. R. Fine tuning model convolutional neural network EfficientNet-B4 dengan augmentasi data untuk klasifikasi penyakit kakao. **Journal of Information System and Application Development**, v. 2, n. 1, p. 1-11, 2024. https://doi.org/10.26905/jisad.v2i1.11899

RECHT, B.; ROELOFS, R.; SCHMIDT, L.; SHANKAR, V. Do ImageNet classifiers generalize to ImageNet? In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 36, 2019, Long Beach. **Proceedings of Machine Learning Research**, v. 97, p. 5389-5400, 2019. Disponible en: https://proceedings.mlr.press/v97/recht19a.html

ROLA, J. B.; BARRERA, J. J. A.; CALHOUN, M. V.; ORAÑO-MAAGHOP, J. F.; UNAJAN, M. C.; BONCALON, J. M.; SEBIOS, E. T.; ESPINOSA, J. S. Convolutional neural network model for cacao *Phytophthora palmivora* disease recognition. **International Journal of Advanced Computer Science and Applications**, v. 15, n. 8, e97, 2024. https://doi.org/10.14569/IJACSA.2024.0150897

SELVARAJU, R. R.; COGSWELL, M.; DAS, A.; VEDANTAM, R.; PARIKH, D.; BATRA, D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION, 2017, Venice. **Proceedings** [...]. Piscataway: IEEE, 2017. p. 618-626. https://doi.org/10.1109/ICCV.2017.74

SUDJUD, S.; JAMIL, M.; MELATI, R. Development of cocoa pod rot disease identification model using You Only Look Once (YOLO-v9). **Mathematical Modelling of Engineering Problems**, v. 12, n. 4, p. 1268-1274, 2025. https://doi.org/10.18280/mmep.120418

SUN, X.; XU, W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. **IEEE Signal Processing Letters**, v. 21, n. 11, p. 1389-1393, 2014. https://doi.org/10.1109/LSP.2014.2337313

TAN, D. S.; LEONG, R. N.; LAGUNA, A. F.; NGO, C. A.; LAO, A.; AMALIN, D. M.; ALVINDIA, D. G. AuToDiDAC: automated tool for disease detection and assessment for cacao black pod rot. **Crop Protection**, v. 103, p. 98-102, 2018. https://doi.org/10.1016/j.cropro.2017.09.017

TAN, M.; LE, Q. EfficientNetV2: smaller models and faster training. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 38, 2021, virtual. **Proceedings of Machine Learning Research**, v. 139, p. 10096-10106, 2021. Disponible en: https://proceedings.mlr.press/v139/tan21a.html

TECHIE-MENSON, H.; ASANTE, M.; MISSAH, Y. M.; ABDUL-SALAAM, G.; OPPONG, S. O. Enhanced convolutional block attention module with learnable gated fusion (LGF-CBAM) for cocoa pod disease identification. **PLOS ONE**, v. 21, n. 4, e0348147, 2026. https://doi.org/10.1371/journal.pone.0348147

TOUVRON, H.; CORD, M.; DOUZE, M.; MASSA, F.; SABLAYROLLES, A.; JEGOU, H. Training data-efficient image transformers & distillation through attention. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 38, 2021, virtual. **Proceedings of Machine Learning Research**, v. 139, p. 10347-10357, 2021. Disponible en: https://proceedings.mlr.press/v139/touvron21a.html

VALERO-GÓMEZ, J. C.; CLARES-PERCA, J. C.; ZÚÑIGA-INCALLA, A. P. **Redes convolucionales y Transformers para clasificar la pudrición negra del cacao (*Phytophthora* spp.): código y datos derivados**. Versión 1.0.1. [S. l.]: Zenodo, 2026. Programa de computador. DOI: https://doi.org/10.5281/zenodo.23178066

VILLALOBOS-CULQUI, C.; GARCÍA-RIVAS-PLATA, C.; TUESTA-HIDALGO, O. A. Artificial vision model based on convolutional neural networks for black pod identification in cacao plantations. **Revista Científica de Sistemas e Informática**, v. 5, n. 1, e678, 2025. https://doi.org/10.51252/rcsi.v5i1.678

WIGHTMAN, R. **PyTorch Image Models**. GitHub, 2019. Repositorio de software. Disponible en: https://github.com/huggingface/pytorch-image-models

YANAC-MONTESINO, R.; ROSALES-HUAMANI, J. A.; CASTILLO-SEQUERA, J. L. Detection of *Phytophthora palmivora* in cocoa fruit with deep learning. In: IBERIAN CONFERENCE ON INFORMATION SYSTEMS AND TECHNOLOGIES (CISTI), 16, 2021, Chaves. **Proceedings** [...]. Piscataway: IEEE, 2021. p. 1-4. https://doi.org/10.23919/CISTI52073.2021.9476279

YOSINSKI, J.; CLUNE, J.; BENGIO, Y.; LIPSON, H. How transferable are features in deep neural networks? In: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 28, 2014, Montreal. **Advances in Neural Information Processing Systems**, v. 27, p. 3320-3328, 2014. Disponible en: https://proceedings.neurips.cc/paper_files/paper/2014/hash/532a2f85b6977104bc93f8580abbb330-Abstract.html

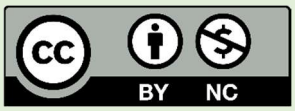
ZECH, J. R.; BADGELEY, M. A.; LIU, M.; COSTA, A. B.; TITANO, J. J.; OERMANN, E. K. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. **PLOS Medicine**, v. 15, n. 11, e1002683, 2018. https://doi.org/10.1371/journal.pmed.1002683

Agradecimientos: Los autores agradecen a la Universidad Nacional de Moquegua por el apoyo institucional, académico y tecnológico brindado durante el desarrollo de esta investigación. Asimismo, se reconoce el acceso a la infraestructura computacional utilizada para la ejecución y validación de los modelos de aprendizaje profundo empleados en el estudio.

Contribuciones de los autores: J.C.V.G.: participó en la conceptualización, desarrollo metodológico, implementación de los modelos, análisis de resultados y redacción del manuscrito; J.C.C.P.: contribuyó en la validación metodológica, revisión del artículo y apoyo en la investigación; A.P.Z.I.: participó en la recopilación y preparación de datos, apoyo técnico y revisión del manuscrito. Todos los autores leyeron y aceptaron la versión publicada del manuscrito.

Disponibilidad de los datos: El conjunto de imágenes es público (PAGADUAN, 2021). El código (notebook, script de entrenamiento y contenedor), las particiones, las predicciones fuera de pliegue, las curvas de entrenamiento y las tablas de resultados están disponibles en https://github.com/juancavalerog/cacao-bpr-cnn-transformers y archivados en Zenodo, versión 1.0.1 (VALERO-GÓMEZ; CLARES-PERCA; ZÚÑIGA-INCALLA, 2026 - https://doi.org/10.5281/zenodo.23178066).

Conflictos de interés: Los autores declaran no tener conflictos de interés. Las entidades financiadoras no participaron en el diseño del estudio; en la recopilación, el análisis o la interpretación de los datos; en la redacción del manuscrito; ni en la decisión de publicar los resultados.



Derechos de autor: © 2026 por los autores. Este artículo es de acceso abierto y se distribuye bajo los términos y condiciones de la licencia Creative Commons Attribution-NonCommercial (CC BY-NC) (https://creativecommons.org/licenses/by/4.0/).