Artículo Científico / Scientific Paper

UNIVERSIDAD POLITÉCNICA SALESIANA ECUADOR
INGENIUS
Revista de Ciencia y Tecnología
https://doi.org/10.17163/ings.n35.2026.10
pISSN: 1390-650X / eISSN: 1390-860X

Swin Transformer V2 para clasificación de café lojano

Swin Transformer V2 for the classification of loja coffee

Patricio Bolívar Betancourt Ludeña[1], Oscar M. Cumbicus Pineda[1,*]

Recibido: 24-07-2025, Recibido tras revisión: 01-12-2025, Aceptado: 10-12-2025, Publicado: 01-01-2026

Resumen

Esta investigación presenta un modelo de clasificación binaria para granos de café verde de la variedad arábico procedentes de la región de Loja, Ecuador, basado en la arquitectura Swin Transformer V2. Se emplearon dos fuentes de datos, el conjunto de datos público USK-Coffee, de origen indonesio, y un conjunto de datos propio capturado bajo condiciones controladas. Se evaluaron dos estrategias de entrenamiento: transferencia secuencial y entrenamiento unificado, siendo este último el que alcanzó una precisión de validación del 98,30 %. Tras la optimización de hiperparámetros, el modelo logró una precisión del 100 % en un conjunto de prueba de 150 imágenes y del 93 % en un conjunto de generalización externo de 400 imágenes con condiciones variables de iluminación y fondo. La interpretabilidad del modelo se validó mediante Grad-CAM, evidenciando que la red enfoca su atención en zonas defectuosas reales. Un análisis de ablación mostró que la disminución de rendimiento en escenarios no controlados se debe principalmente a la sensibilidad al ruido y a la iluminación extrema. Como principales aportes, se destaca la creación de un conjunto de datos especializado y un modelo eficiente para la clasificación automática de café verde arábico.

Palabras clave: clasificación de café, inteligencia artificial, Vision Transformer, Swin Transformer, visión por computadora, aprendizaje por transferencia

Abstract

This study presents a binary classification model for green coffee beans of the Arabica variety from the Loja region in Ecuador, based on the Swin Transformer V2 architecture. Two datasets were used, the public USK-COFFEE dataset of Indonesian origin and a proprietary dataset captured under controlled conditions. Two training strategies were evaluated: sequential transfer learning and unified training, with the latter achieving a validation accuracy of 98.30%. After hyperparameter optimization, the model reached 100% accuracy on a test set of 150 images and 93% accuracy on an external generalization set of 400 images with varying lighting conditions and backgrounds. Model interpretability was validated using Grad-CAM, demonstrating that the network focuses on actual defective regions rather than background information. An ablation analysis revealed that performance degradation in unconstrained scenarios is mainly due to sensitivity to noise and extreme lighting conditions. The main contributions of this work include the creation of a specialized dataset for Arabica green coffee from Loja and the development of an efficient model for its automatic classification.

Keywords: coffee classification, artificial intelligence, Vision Transformer, Swin Transformer, computer vision, transfer learning

[1,*]Carrera de Ingeniería en Computación, Universidad Nacional de Loja, Ecuador.
Autor para correspondencia ✉: oscar.cumbicus@unl.edu.ec.

Forma sugerida de citación: P. T. Betancourt Ludeña y O. M. Cumbicus Pineda, "Swin Transformer V2 para clasificación de café lojano," *Ingenius, Revista de Ciencia y Tecnología*, N.° 35, pp. 138-148, 2026. DOI: https://doi.org/10.17163/ings.n35.2026.10.

1. Introducción

El sector de producción de café, específicamente en la producción de café verde arábico, está experimentando un crecimiento continuo [1]. De la misma forma, en la zona de Loja, Ecuador, la expansión de este sector ha sido evidente [2]. Sin embargo, enfrenta retos significativos en cuanto a la clasificación precisa de los granos, un proceso esencial para asegurar la calidad del producto final y mantener su competitividad en el mercado internacional [3]. Tradicionalmente, este proceso se ha realizado de manera manual, lo que ha dado lugar a variabilidad en los resultados debido a la intervención humana y a la fatiga del trabajo repetitivo [4]. En este contexto, las tecnologías que han aparecido en los últimos años, basadas en inteligencia artificial (IA), específicamente en modelos de visión por computadora, han ofrecido una solución innovadora y precisa para la automatización de la tarea de clasificación de granos de café [5, 6]. Hasta el momento, las redes neuronales convolucionales (CNN) han destacado por demostrar resultados muy favorables en la clasificación [7, 8], [5]. Por otro lado, los modelos Vision Transformers (ViT), y más recientemente el Swin Transformer, han demostrado una superioridad al generalizar ciertas características locales y globales de imágenes complejas [9, 10].

En la actualidad, son variados los modelos que se presentan como soluciones a esta problemática, como en [11] donde se alcanzó una precisión de 99.84 % en la detección de defectos de granos de café verde arábico. En el trabajo realizado en [12], se emplearon imágenes multiespectrales y SVM (Support Vector Machine, máquina de vectores de soporte) para clasificar entre clases especiales y comerciales de granos de café, alcanzando 96 % de precisión. Asimismo, en [13] se usó un enfoque de clasificación multiclase y se reportó una precisión de 84.75 % con su modelo Swin Transformer, al usar el conjunto de datos USK-Coffee. Estos trabajos ilustran la capacidad que tienen las tecnologías de visión por computadora para superar las limitaciones humanas en la clasificación de granos de café. De la misma manera, una limitación importante para el desarrollo de los modelos a nivel local es la inexistencia de un conjunto de datos propio de la zona de Loja, lo cual restringe la capacidad de los modelos para generalizar y aprender las características específicas del café de producción lojano.

Recientemente, los modelos de visión por computadora basados en Transformers han demostrado un gran potencial [14, 15]. Los Vision Transformers procesan las imágenes en parches, aplicando una atención global y local a los detalles de estas. En trabajos comparativos como en [16] donde se compararon modelos de ViT y CNN en imágenes de retina, se demostró que el modelo Swin Transformer obtuvo un rendimiento mayor que las redes neuronales convolucionales con una precisión del 97.3 %. Adicionalmente, el modelo Swin Transformer V2 desarrollado por Liu *et al.* [17] ha escalado su arquitectura a 3 mil millones de parámetros y estableció récords de rendimiento al aumentar la capacidad y resolución del modelo. Este avance en la versión del modelo sugiere una mayor capacidad para la discriminación de patrones visuales más complejos, lo cual es necesario para la detección de defectos pequeños en los granos de café. El presente trabajo propone el diseño y evaluación de un modelo de clasificación binaria de granos de café verde arábico de la zona de Loja, Ecuador.

En consecuencia, para la realización de este estudio se emplearon conjuntos de datos públicos de imágenes de café, complementados con un conjunto de imágenes tomadas de granos de la zona de Loja, obtenidos mediante una captura en condiciones controladas. El modelo fue entrenado y validado con los conjuntos de datos antes descritos y evaluado con el conjunto de datos lojano. Este estudio evaluó el rendimiento del modelo de Swin Transformer V2, preentrenado con ImageNet-1K, a través de la aplicación de técnicas de transfer learning y fine tuning. Por otro lado, esta investigación contribuye al estado del arte un nuevo conjunto de datos de café verde arábico de la provincia de Loja.

2. Materiales y métodos

En la presente sección se describen los conjuntos de datos utilizados, la arquitectura del modelo Swin Transformer V2 implementado, la estrategia de entrenamiento y las métricas de evaluación empleadas para la clasificación de granos de café verde arábico, todo lo anterior resumido en 3 fases abarcadas por la metodología CRISP-ML(Q) [18], como lo son ingeniería de datos, ingeniería de modelos y evaluación de modelos (Figura 1).

2.1. Conjuntos de datos utilizados

Para el desarrollo y evaluación del modelo, se utilizaron dos conjuntos de datos principales, relacionados con la clasificación de granos de café verde (pelado y sin someter a tueste) de la variedad arábica. A continuación, se pueden ver los conjuntos de datos seleccionados para esta investigación.

USK-Coffee: Un conjunto de datos público que contiene 8000 imágenes de granos de café verde arábico, originalmente distribuidas en 4 clases (peaberry, longberry, premium, defect) [19]. Para este estudio, se modificó el conjunto de datos original consolidando tres de sus clases (peaberry, longberry y premium) en una categoría denominada “buenos”, mientras que la clase defect se mantuvo como “defectuosos”, formando así una clasificación binaria. Específicamente, a la clase “defectuosos” se le aplicaron técnicas de aumento de

datos, incluyendo: rotaciones aleatorias, volteos horizontales/verticales, ajustes de brillo/contraste, con el fin de balancear las clases manteniendo cierta variabilidad de los datos, para evitar un overfitting en el entrenamiento.

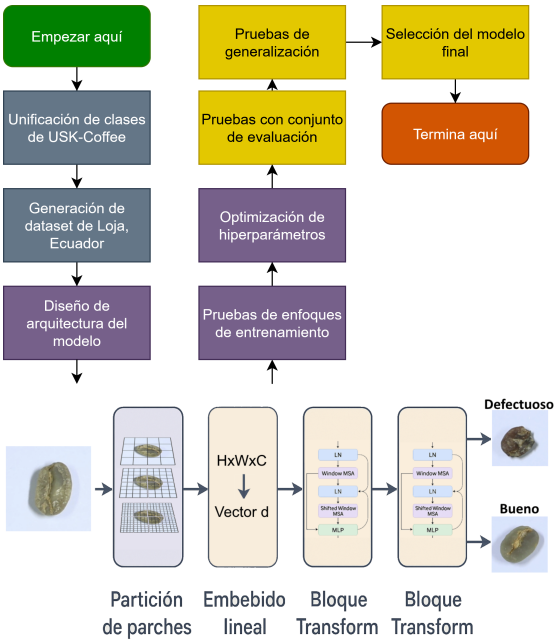


Figura 1. Diagrama de materiales y métodos. El flujo muestra: entrada de imagen de grano de café (HxWxC), partición en parches, embebido lineal, bloques Transformer secuenciales y clasificación binaria final (bueno/defectuoso).

Conjunto de datos lojano. Se construyó un conjunto de datos propio a partir de granos de café verde arábico recolectados en la zona de Gonzanamá, provincia de Loja, Ecuador. Para la recolección, se seleccionaron cinco libras de café de manera aleatoria, las cuales fueron descascaradas y trilladas. Posteriormente, un caficultor local realizó la clasificación manual separando los granos en dos categorías: bueno y defectuoso. Con el fin de capturar las imágenes bajo condiciones controladas, se dispusieron los granos sobre una hoja A4 de papel bond blanco colocada encima de una cartulina A3 blanca. Cada fotografía incluyó 50 granos organizados en 5 columnas y 10 filas, con separaciones de 5 cm y 3 cm, respectivamente. La iluminación se aseguró mediante dos lámparas LED blancas colocadas a 22 cm de altura y una tira LED RGB en blanco alrededor de la hoja.

Las imágenes fueron tomadas con una cámara Canon EOS R50 equipada con un lente de 20 mm, ubicada a 28 cm de la superficie. Los parámetros de captura se configuraron en: tiempo de obturación de 1/5 s, apertura F22, sensibilidad ISO 100 y disparo remoto con temporizador de 2 segundos. Cada fotografía se registró en formatos JPG y RAW, con resolución de 6000×4000 px (24 MP).

En total, se capturaron 10 fotografías de 50 granos cada una, las cuales fueron procesadas posteriormente mediante un script en Python para recortar imágenes individuales de 256×256 px en formato RGB, manteniendo un fondo blanco. El resultado final fue un conjunto balanceado de 1000 imágenes (500 por clase) que conforman el conjunto de datos lojano.

Ambos conjuntos de datos fueron divididos aleatoriamente en subconjuntos de entrenamiento, validación y prueba. Para el conjunto de datos USK-Coffee modificado, se utilizó una proporción de 80 % para entrenamiento y 20 % para validación. Para el conjunto de datos lojano, la división fue de 70 % para entrenamiento, 15 % para validación y 15 % para prueba (Figura 2).

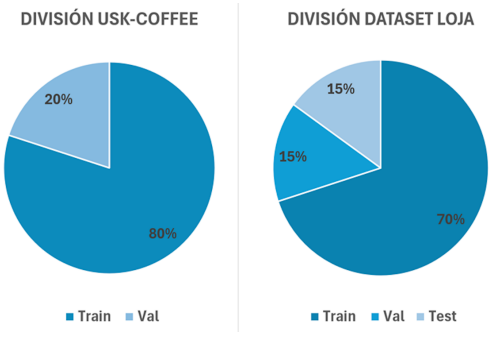


Figura 2. Porcentajes de división de los conjuntos de datos USK-Coffee y lojano.

Para fomentar la reproducibilidad y la colaboración en la comunidad científica, el conjunto de datos lojano se ha puesto a disposición pública en el repositorio Zenodo bajo una licencia CC BY-NC 4.0 (Creative Commons Attribution No Derivatives 4.0 International), permitiendo su uso en investigación sin fines de lucro. El conjunto de datos puede ser accedido a través de su DOI [20].

2.2. Modelo Swin Transformer V2

El modelo de aprendizaje profundo seleccionado fue un Swin Transformer V2. Esta arquitectura fue elegida por su capacidad de extraer características tanto locales como globales de las imágenes, gracias a su diseño jerárquico con mecanismos de autoatención en ventanas desplazadas que permiten captar patrones a diferentes escalas. Se exploraron tres variantes preentrenadas del Swin Transformer V2 (denominadas base, large y giant), las cuales provienen de un entrenamiento previo en el conjunto de datos ImageNet-1K [17].

La variante seleccionada sirvió como punto de partida para realizar fine tuning en la tarea específica de clasificación de granos de café lojano, ofreciendo diferentes capacidades de modelo para ser comparadas en los experimentos. En general, el uso de pesos preentrenados aprovechó el conocimiento adquirido en la clasificación general de imágenes, lo que aceleró la convergencia del entrenamiento en el nuevo dominio. La elección de esta arquitectura se decidió por las peculiaridades de Transformers en visión por computadora, destacando en su capacidad para capturar dependencias globales en la imagen mediante mecanismos de atención, a diferencia de las CNN tradicionales enfocadas en regiones locales.

2.3. Métodos de entrenamiento

Se evaluaron dos métodos de entrenamiento del modelo con los datos seleccionados. El primer método correspondió a una aplicación secuencial de transfer learning, en la cual se ajustó el modelo en dos etapas, en la primera, se realizó el fine tuning inicial utilizando únicamente el conjunto de datos USK-Coffee, y posteriormente se llevó a cabo una segunda etapa de ajuste fino empleando las imágenes locales del café de Loja; generalizando las características del conjunto amplio antes de especializarse en las características del conjunto reducido. El segundo método consistió en un entrenamiento conjunto, combinando los dos conjuntos de datos desde el inicio en una única fase de entrenamiento. En este método, las imágenes de USK-Coffee y de Loja se mezclaron durante el proceso de aprendizaje, de manera que el modelo aprendió simultáneamente de ambos conjuntos de datos. Estos dos enfoques permitieron comparar los resultados de entrenar el modelo con el conjunto de datos global de manera secuencial e integrarlo de forma unificada durante la optimización del modelo.

2.4. Configuración experimental

Los experimentos se llevaron a cabo utilizando recursos de cómputo con soporte de GPU para acelerar el entrenamiento. En particular, se empleó una estación de trabajo local (entorno Visual Studio Code) equipada con GPU de 8 GB de VRAM y un servidor institucional con JupyterHub con capacidad de cómputo gráfico de 4 GPUs con 12 GB de VRAM cada una (48 GB de VRAM total).

En cuanto a la optimización de hiperparámetros de entrenamiento, se generaron 7 modelos, probando diferentes valores de tasa de aprendizaje, tamaño de lote y dropout (Tabla 1). El número de épocas de entrenamiento se fijó de antemano con base en experimentos preliminares [18], dando las iteraciones suficientes para la convergencia del modelo sin llegar a un sobreajuste.

Tabla 1. Valores aplicados en la optimización de hiperparámetros manual. Ép: Épocas, TL: Tamaño de lote (batch size), TA: Tasa de aprendizaje (learning rate), Opt.: Optimizador, Acr.: Acrónimo del modelo.

Ép	TL	TA	Dropout	Opt.	Acr.
25	2	3×10^{-6}	0.5	Adam	M1
25	2	1×10^{-6}	0.5	Adam	M2
20	2	5×10^{-6}	0.5	Adam	M3
20	2	7×10^{-6}	0.45	Adam	M4
40	4	3×10^{-6}	0.55	Adam	M5
100	4	5×10^{-7}	0.55	Adam	M6
100	4	5×10^{-7}	0.55	SGD	M7

En términos de complejidad computacional, el entrenamiento del modelo final (configuración M4, ver Tabla 1) requirió aproximadamente una hora utilizando el servidor institucional. Este tiempo de entrenamiento relativamente corto, dado el tamaño del conjunto de datos combinado, demuestra la eficiencia del enfoque de transfer learning. Para una implementación práctica en un entorno de producción, como una planta de procesamiento de café, se requeriría una estación de trabajo con una GPU dedicada de gama media o alta para asegurar una clasificación en tiempo real. Aunque esto implica una inversión inicial en hardware, el costo es competitivo en comparación con equipos de selección óptica especializados, ofreciendo una solución escalable y adaptable mediante software.

3. Resultados y discusión

3.1. Ingeniería de datos

Se consolidaron los conjuntos de datos necesarios y se aplicaron técnicas para mejorar su calidad. Inicialmente, se seleccionó el conjunto de datos USK-Coffee

como base debido a su amplitud (8000 imágenes de café verde arábico). Dado que la tarea se planteó como una clasificación binaria de granos "buenos" y "defectuosos", las clases originales (peaberry, longberry y premium) se unificaron en una sola categoría denominada "bueno" y por su parte la clase "defect" se renombró como "defectuoso" (como se describe en la sección 2.1). Este conjunto de datos modificado dio como resultado un desbalance de 3000 imágenes de granos buenos, frente a 2000 imágenes de granos defectuosos. Para la corrección del desequilibrio de clases, se realizó un sobremuestreo con transformaciones (Figura 3), de la clase defectuoso hasta obtener 3000 imágenes por cada clase.

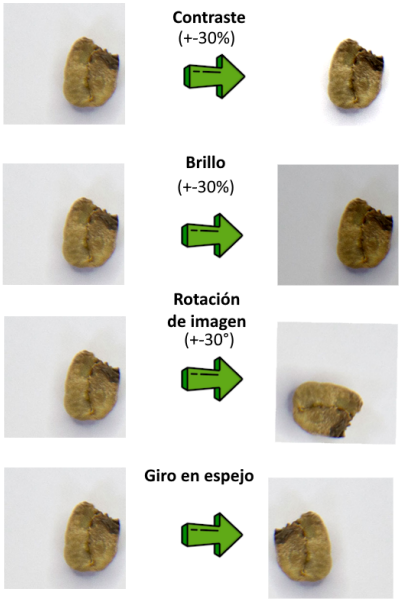


Figura 3. Transformaciones para sobremuestreo de la clase minoritaria "defectuoso".

Adicionalmente, se construyó un conjunto de datos de café verde arábico recolectado y clasificado por productores de Loja. Se capturaron 500 imágenes por clase ("bueno" y "defectuoso"), aportando un contexto local con las características intrínsecas del café de la zona, complementando el conjunto de datos de Indonesia. Los conjuntos de datos de Indonesia y Loja se dividieron aleatoriamente para el entrenamiento y validación del modelo, y un 15 % del conjunto de datos de Loja se destinó para la prueba del desempeño del modelo, como se describe en la sección 2.1.

3.2. Ingeniería de modelos

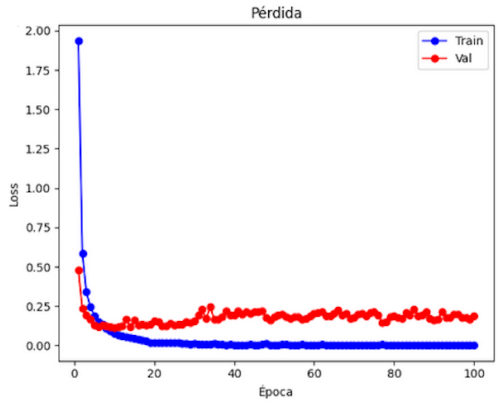
Se decidió emplear la arquitectura Swin Transformer V2 preentrenada, ajustándola mediante fine tuning para la tarea específica de clasificación de granos de café verde arábico. Entre las variantes disponibles, se seleccionó el modelo Swin V2 Large debido a que ofrece un equilibrio adecuado entre rendimiento y complejidad: reporta un Top-1 Accuracy de 87.7 % en ImageNet [17] con un tamaño moderado de parámetros, evitando la sobrecarga computacional de la versión gigante (90.2 % Top-1 pero mucho más pesada).

Se exploraron dos métodos de entrenamiento del modelo para aprovechar los datos disponibles. En el Método 1, se entrenó primero el modelo con él conjunto de datos USK-COFFEE (Indonesia) y luego se realizó un transfer learning adicional con el conjunto de datos de Loja, buscando que en una segunda fase el modelo se especialice en las características regionales. En el Método 2, en cambio, se combinaron ambos conjuntos de datos desde el inicio en un solo entrenamiento unificado. Tras probar ambos enfoques, se observó que el método de conjunto de datos unificado logró una mayor precisión de validación (98.30 %) que el método por fases separado (97.33 %). Por ello, para las siguientes etapas se optó por entrenar el modelo con los datos combinados, lo que sugiere que exponer conjuntamente al modelo al completo de datos permitió una generalización inicial mejor.

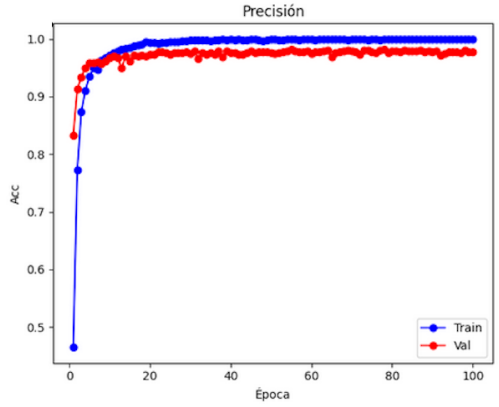
Adicionalmente, se llevó a cabo una optimización manual de hiperparámetros, ajustando iterativamente los valores hasta maximizar la precisión de validación, para refinar el desempeño del modelo. Se probaron 7 configuraciones distintas (modelos M1–M7) variando parámetros clave como número de épocas, tamaño de lote (batch), tasa de aprendizaje (learning rate), dropout y optimizador. Cada modelo fue entrenado bajo la estrategia de datos unificados. En general, todos los modelos lograron converger a altas precisiones de validación sin diferencias drásticas entre la mayoría de las configuraciones. Este resultado indica que la elección de hiperparámetros dentro de rangos razonables no afectó radicalmente el desempeño, debido a la riqueza del conjunto de datos y la robustez del modelo preentrenado. No obstante, para proceder con una evaluación más detallada, se seleccionaron los cuatro modelos con mayor desempeño (M1, M4, M5 y M6), siendo M6 el que mejor desempeño demostró en el entrenamiento y validación con 98.22 % de precisión global (Figura 4); además, se descartaron los que presentaron un menor rendimiento (Tabla 2).

Tabla 2. Valores de precisión y pérdida en validación para cada modelo en la optimización de hiperparámetros.

Modelo	Precisión validación	Pérdida validación
M1	98.15 %	0.1242
M2	97.85 %	0.1009
M3	97.85 %	0.1401
M4	97.93 %	0.1031
M5	97.93 %	0.1749
M6	98.22 %	0.1527
M7	96.67 %	0.1235



(a)



(b)

Figura 4. Progreso en el entrenamiento y validación del modelo M6. a) Curva de pérdida, b) Curva de precisión.

3.3. Evaluación de los modelos

Se analizó el rendimiento de los modelos seleccionados en el conjunto de prueba (test), tanto en un entorno controlado (conjunto de pruebas) como en condiciones más desafiantes, integrando las métricas de clasificación y su interpretación a la luz del marco teórico. Inicialmente, los cuatro modelos (M1, M4, M5, M6) fueron evaluados sobre el conjunto de prueba local (imágenes de Loja no utilizadas en entrenamiento, 150 muestras). Todos lograron una precisión global del 100 % en esta prueba, clasificando correctamente cada imagen como grano "bueno" o "defectuoso" (Figura 5). Consecuentemente, sus métricas por clase (precisión, *recall* –sensibilidad– y puntaje F1 media armónica entre precisión y *recall*) alcanzaron valores de 100 % en ambas categorías, evidenciando una adaptación sobresaliente a las características particulares de los granos de café lojano. Si bien un 100 % de acierto sugiere un desempeño excelente, es importante considerar que este conjunto de prueba proviene del mismo dominio y condiciones controladas que el entrenamiento (mismas variedades de grano, iluminación similar, fondo uniforme). Desde una perspectiva crítica, un resultado perfecto en un entorno conocido puede indicar un sobreajuste del modelo a las condiciones específicas, por lo que era necesaria una prueba más exigente para evaluar la capacidad de generalización de los modelos.

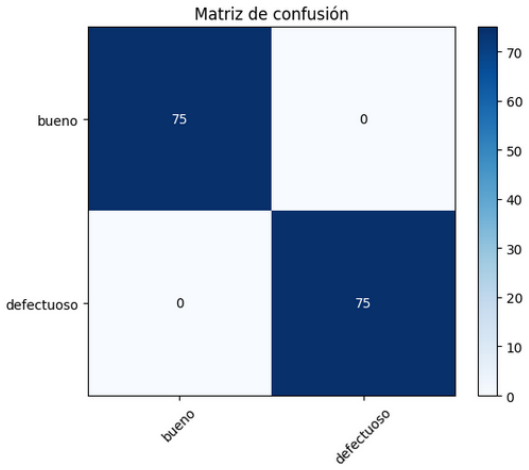


Figura 5. Matriz de confusión de los modelos con el conjunto de prueba.

Prueba de generalización. Para una evaluación más rigurosa basada en la metodología CRISP-ML(Q), se sometió al modelo a un conjunto de pruebas de generalización compuesto por 400 imágenes adicionales con mayores variaciones (condiciones de luz, fondo, dispositivo de captura, entre otras), simulando un entorno real. Estas imágenes fueron capturadas con dispositivos móviles bajo condiciones no controladas para evaluar la robustez del modelo. En este escenario, se evidenció una degradación marginal del desempeño; no obstante, el rendimiento global del modelo se mantuvo elevado, el mejor de los modelos (configuración M4, ver Tabla 1) alcanzó una precisión global del 93 %, mientras que los demás rondaron el 92 %. La Figura 6 muestra la matriz de confusión obtenida por el modelo M4 en esta prueba de generalización. En ella se observa que, de 200 imágenes reales de cada clase, el modelo identificó correctamente 195 granos "buenos" y 176 "defectuosos".

Se produjeron algunos errores con 5 falsos negativos y 24 falsos positivos. Estas cifras implican que, para la clase "bueno", se logró una sensibilidad (*recall*) del 97 % y un *F1-score* (media armónica entre precisión y *recall*) del 93 %, mientras que para la clase "defectuoso" la sensibilidad fue menor con 88 % y un *F1-score* de 92 % debido principalmente a los falsos positivos. A pesar de esta leve disminución de rendimiento en condiciones más heterogéneas, el modelo mantiene un equilibrio aceptable entre precisión y *recall* en ambas clases, evitando sesgarse completamente hacia una de ellas. Esto sugiere que la estrategia de balanceo de datos y la incorporación de variabilidad en el entrenamiento efectivamente contribuyeron a lograr una clasificación consistente en ambas categorías de granos.

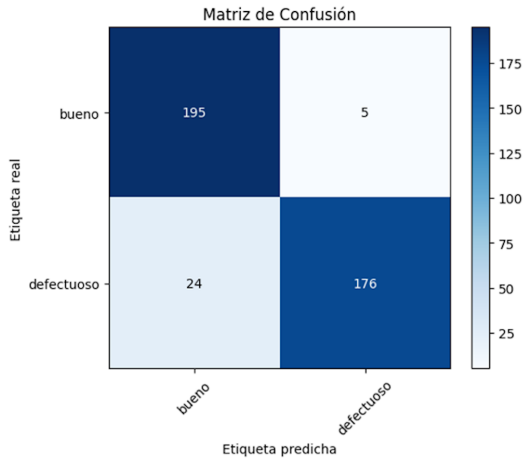


Figura 6. Matriz de confusión de la prueba de generalización del modelo con la cuarta configuración (M4).

Un análisis más detallado de los errores en la prueba de generalización ofrece información valiosa. Los 24 falsos positivos se atribuyen principalmente a las condiciones de captura no controladas. La iluminación variable en las imágenes de prueba, tomadas con dispositivos móviles bajo condiciones no controladas (a diferencia del conjunto de entrenamiento), probablemente generó sombras que ocultaron defectos sutiles, llevando al modelo a una clasificación incorrecta. Por otro lado, los 5 falsos negativos pueden explicarse por factores similares; sombras pronunciadas o manchas naturales en granos sanos pudieron ser malinterpretadas como defectos por el modelo. Adicionalmente, la menor resolución de las cámaras de los dispositivos móviles pudo afectar la claridad de los detalles del grano, contribuyendo a estas confusiones.

Para complementar estas métricas, el reporte de clasificación del modelo M4 arrojó una precisión balanceada (macro avg –promedio de las métricas de cada clase–) del 93 %. Este valor, junto con los altos puntajes de *recall* (97 % para "bueno" y 88 % para "defectuoso"), confirma que el modelo mantiene un rendimiento equilibrado y no muestra un sesgo significativo hacia ninguna de las dos clases, abordando una de las preocupaciones clave en tareas de clasificación con clases desbalanceadas.

Finalmente, aunque el modelo M6 obtuvo la mayor precisión en la validación, el modelo M4 demostró una mayor capacidad de generalización en un conjunto de datos más desafiante y no visto, por presentar una precisión global superior (93 %), superando al resto de los modelos (Tabla 3). además, mostró métricas por clase más uniformes, a diferencia de los otros modelos que, aunque tienen una precisión global similar, evidenciaron pequeñas brechas entre clases. El balance entre clases resulta fundamental en tareas de clasificación binaria desbalanceada, ya que evidencia que el modelo no prioriza la detección de una categoría sobre la otra, sino que mantiene una capacidad de reconocimiento equilibrada y precisa para ambos tipos de granos. Los hallazgos de la evaluación indican que el modelo Swin Transformer V2 ajustado puede alcanzar resultados perfectos bajo condiciones controladas; no obstante, en contextos más heterogéneos su precisión puede disminuir, lo que sugiere oportunidades de mejora en la robustez del enfoque implementado.

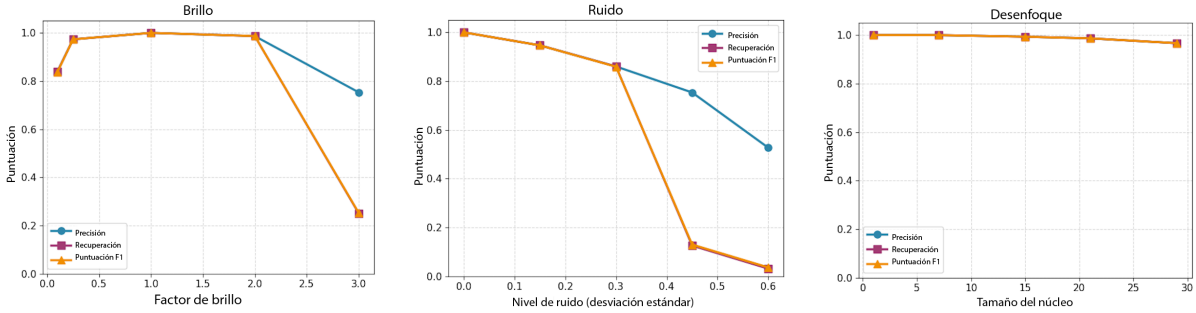
Tabla 3. Valores de precisión y pérdida en validación para cada modelo en la optimización de hiperparámetros.

Modelo	Precisión	Precisión 'bueno'	Precisión 'defectuoso'
M1	0.92	0.88	0.96
M4	**0.93**	**0.89**	**0.97**
M5	0.92	0.89	0.96
M6	0.82	0.88	0.96

Análisis de ablación. Adicionalmente, para explicar la divergencia de rendimiento entre el entorno controlado (100 %) y la prueba de generalización (93 %), se realizó un estudio de ablación cuantitativo sometiendo al modelo a degradaciones ambientales sintéticas: variaciones de brillo, ruido y desenfoque (Figura 7). El análisis revela que la arquitectura es altamente robusta al desenfoque (Blur), manteniendo un *F1-score* superior al 96 % incluso con núcleos de desenfoque grandes (Figura 7c), lo que indica que la falta de nitidez en fotos de móviles no es un factor crítico. Sin embargo, el modelo mostró sensibilidad ante condiciones extremas de iluminación y ruido. Como se observa en la Figura 7a (Brightness), aunque el modelo tolera variaciones moderadas, una sobreexposición severa (factor > 2.5) provoca una caída drástica del *recall* al 25 %, impidiendo la detección de defectos. De igual forma, la inyección de ruido (noise, Figura 7b) degrada linealmente la precisión; con una desviación estándar de 0.6, el *recall* desciende a un 3 %, sugiriendo que el grano digital generado por sensores móviles en baja luz afecta la capacidad del modelo para discernir texturas finas. Estos hallazgos confirman que la caída del 7 % en la prueba de generalización se atribuye principalmente a factores de iluminación no controlada y ruido de sensor, más que a problemas de enfoque.

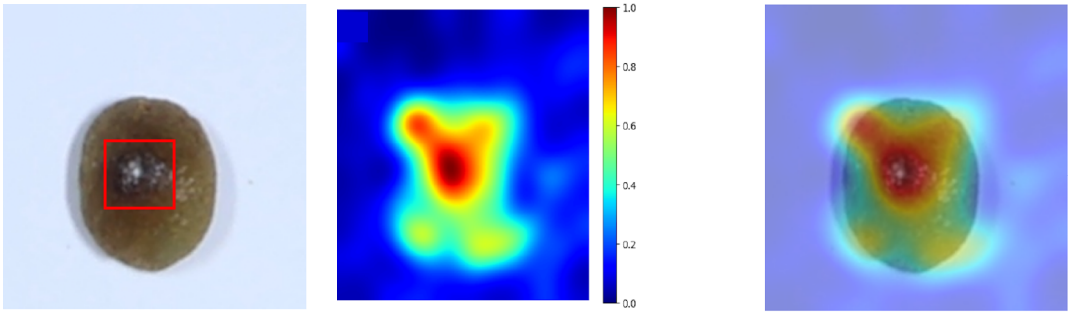
Pruebas de interpretabilidad. Para complementar el análisis cuantitativo y asegurar la interpretabilidad del modelo, se empleó la técnica de visualización Grad-CAM [21]. Esta herramienta permite verificar si la red neuronal fundamenta sus predicciones en las características relevantes del grano o en artefactos espurios. En la Figura 8a se presenta la imagen original de un grano clasificado como defectuoso, donde se ha delimitado mediante un recuadro rojo la ubicación física real del defecto antes del procesamiento. Los resultados visuales posteriores, mostrados en el mapa de calor (Figura 8b) y la superposición Grad-CAM (Figura 8c), evidencian que el modelo Swin Transformer concentra su máxima activación (representada en tonos rojos)

precisamente dentro del área del defecto señalada. Esta coincidencia espacial entre el defecto real y la atención de la red valida empíricamente el proceso de toma de decisiones del modelo, confirmando que la clasificación se basa en la morfología del daño y no en el ruido de fondo.



(a) (b) (c)

Figura 7. Análisis de ablación frente a perturbaciones ambientales. a) Impacto de la variación de brillo (brightness). b) Impacto del ruido gaussiano (noise). c) Impacto del desenfoque (blur). Se observa una alta estabilidad ante el desenfoque, pero una degradación significativa del *recall* en condiciones de ruido extremo y sobreexposición.



(a) Original (b) Mapa de calor (c) Grad-CAM Pred: Defectuoso (91,2 %)

Figura 8. Análisis de interpretabilidad mediante Grad- CAM. a) Imagen original con el defecto resaltado manualmente en un recuadro rojo. b) Mapa de calor de atención generado por el modelo. c) Superposición (Grad-CAM) mostrando que la activación del modelo coincide con la zona del defecto.

3.4. Discusión

Los resultados obtenidos se contrastan a continuación con estudios recientes sobre clasificación de granos de café verde mediante inteligencia artificial, especialmente aquellos basados en arquitecturas Vision Transformer y CNN, para contextualizar el rendimiento de nuestro enfoque para café de Loja.

En primer lugar, el modelo Swin V2 adaptado alcanzó en validación interna precisiones de 98 % y en pruebas finales de generalización de hasta 93 % de acierto global. Este desempeño se alinea con la tendencia reportada por trabajos recientes que exploran Transformers en visión, como en [13] se logró un 84.75 % de precisión utilizando una arquitectura basada en Transformer (Swin Transformer) en una clasificación multiclase con el mismo conjunto de datos USK-Coffee sin modificaciones. En el presente estudio, al simplificar el problema a una clasificación binaria y reforzarlo con datos locales y fine tuning, se alcanzó durante la fase de validación un rendimiento de 98 % de precisión global. Del mismo modo, en [15] se propuso un método de clasificación de granos con Swin Transformer orientado a la gradación de calidad, destacando igualmente el potencial de esta arquitectura para capturar características relevantes en café verde. Los hallazgos obtenidos respaldan esas evidencias: los Vision Transformers, adecuadamente entrenados, pueden alcanzar una precisión muy elevada en esta tarea, incluso por encima de algunos modelos CNN tradicionales bajo escenarios similares.

En entornos de prueba controlados, el modelo ViT entrenado en el presente estudio consiguió un 100 % de acierto, superando las precisiones reportadas por CNN en tareas similares. En [6] se diseñó una CNN optimizada para detección de defectos en granos de café que alcanzó 95.2 % de precisión en su conjunto de

prueba. En contraste, el presente modelo obtuvo 100 % en pruebas locales y 93 % en pruebas de mayor complejidad, demostrando la capacidad de los ViT para lograr un desempeño similar o superior a las CNN en la clasificación de granos. Sin embargo, es necesario contextualizar estos números. En [5] se desarrolló una CNN ligera y explicable (LDCNN) para detección de calidad de café verde, logrando 98.38 % de precisión y 98.24 % de *F1-score* en su conjunto de datos de prueba. Asimismo, Gope y Fukai [22] reportaron cerca de 98.19 % de precisión usando una CNN para clasificar granos tipo "peaberry" *vs* "normales". Estas cifras superan ligeramente el 93 % obtenido por el modelo en la prueba de generalización; no obstante, cabe resaltar que estos estudios realizaron pruebas en entornos controlados o con condiciones fijas.

Por otro lado, en [11] emplearon una arquitectura MobileNetV3 para clasificar defectos en café arábigo tailandés (tarea multiclase) y obtuvieron 88.63 % de precisión, valor inferior al 93 % alcanzado por el modelo entrenado en un reto de generalización más complejo. En conjunto, estos contrastes sugieren que el enfoque usado en este trabajo basado en ViT, es similar con el estado del arte recabado, logrando resultados de métricas favorables en contraste a muchos de los modelos previos de CNN, especialmente al simplificar la clasificación a dos clases y al sumar un nuevo conjunto de datos especializado en la zona de Loja. A su vez, el análisis de robustez revela que, aunque el modelo supera a arquitecturas previas, existe un margen de mejora para equiparar la tolerancia al ruido que algunas CNN han demostrado, sugiriendo la necesidad de incluir aumentos de datos específicos de ruido e iluminación en trabajos futuros.

4. Conclusiones

La creación y evaluación del modelo de Vision Transformer para la clasificación de granos de café verde arábico lojano ha demostrado una capacidad de discriminación alta, alcanzando una precisión global de 100 % en condiciones controladas y de 93 % en pruebas de generalización con condiciones variadas.

La adopción de la metodología CRISP-ML(Q), acompañada de la comparación inicial de los métodos de entrenamiento con los conjuntos de datos USK-Coffee y el propio de Loja, permitió una convergencia más rápida y estable del modelo. Asimismo, la optimización manual de hiperparámetros: épocas, tamaño de lote (batch size), tasa de aprendizaje (learning rate), dropout y optimizador resultó clave para refinar su desempeño, logrando precisiones de validación superiores al 98 % sin grandes oscilaciones entre las configuraciones evaluadas.

Estos hallazgos confirman el potencial de los Vision Transformers como herramienta de control de calidad en la industria cafetalera de Loja. No obstante, la ligera disminución de precisión bajo condiciones menos ideales demuestra la existencia de un margen de mejora, donde se puede ampliar la diversidad del conjunto de datos, incorporar nuevas técnicas de aumento de datos y explorar esquemas de optimización de hiperparámetros. En consecuencia, esto podría fortalecer la capacidad de generalización y asegurar resultados igualmente robustos en escenarios menos controlados.

Agradecimientos

Extendemos nuestro agradecimiento a la Universidad Nacional de Loja y a la Carrera de Ingeniería en Computación por el apoyo y los recursos facilitados para este trabajo.

Rol de autores

- **Patricio Bolívar Betancourt Ludeña:** escritura – borrador original, diseño de software y curación de datos.
- **Oscar M. Cumbicus Pineda:** supervisión y validación.

Referencias

[1] ICP. (2025) I-CIP retreats on news of looser supply, relieving some of the upward pressure. Coffee market report. International Coffee Organization. [Online]. Available: https://upsalesiana.ec/ing35ar10r1

[2] Agricultura. (2025) 6425 hectáreas de café son renovadas en la provincia de Loja. Ministerio de Agricultura, Ganadería y Pesca. [Online]. Available: https://upsalesiana.ec/ing35ar10r2

[3] M. Faisal, J.-S. Leu, and J. T. Darmawan, "Model selection of hybrid feature fusion for coffee leaf disease classification," *IEEE Access*, vol. 11, pp. 62 281–62 291, 2023. [Online]. Available: https://doi.org/10.1109/ACCESS.2023.3286935

[4] E. Hassan, "Enhancing coffee bean classification: a comparative analysis of pretrained deep learning models," *Neural Computing and Applications*, vol. 36, no. 16, pp. 9023–9052, Apr. 2024. [Online]. Available: https://doi.org/10.1007/s00521-024-09623-z

[5] C.-H. Hsia, Y.-H. Lee, and C.-F. Lai, "An explainable and lightweight deep convolutional neural network for quality detection of green coffee beans," *Applied Sciences*, vol. 12, no. 21, p. 10966, Oct. 2022. [Online]. Available: https://doi.org/10.3390/app122110966

[6] S.-J. Chang and C.-Y. Huang, "Deep learning model for the inspection of coffee bean defects," *Applied Sciences*, vol. 11, no. 17, p. 8226, Sep. 2021. [Online]. Available: https://doi.org/10.3390/app11178226

[7] A. Chavarro, D. Renza, and E. Moya-Albor, "Convnext as a basis for interpretability in coffee leaf rust classification," *Mathematics*, vol. 12, no. 17, p. 2668, Aug. 2024. [Online]. Available: https://doi.org/10.3390/math12172668

[8] Y. A. Auliya, I. Fadah, Y. Baihaqi, and I. N. Awwaliyah, "Green bean classification: Fully convolutional neural network with Adam optimization," *Mathematical Modelling of Engineering Problems*, vol. 11, no. 6, pp. 1641–1648, Jun. 2024. [Online]. Available: https://doi.org/10.18280/mmep.110626

[9] J. Maurício, I. Domingues, and J. Bernardino, "Comparing vision transformers and convolutional neural networks for image classification: A literature review," *Applied Sciences*, vol. 13, no. 9, p. 5521, Apr. 2023. [Online]. Available: https://doi.org/10.3390/app13095521

[10] J. Wei, J. Chen, Y. Wang, H. Luo, and W. Li, "Improved deep learning image classification algorithm based on Swin Transformer V2," *PeerJ Computer Science*, vol. 9, p. e1665, Oct. 2023. [Online]. Available: https://doi.org/10.7717/peerj-cs.1665

[11] S. Arwatchananukul, D. Xu, P. Charoenkwan, S. Aung Moon, and R. Saengrayap, "Implementing a deep learning model for defect classification in Thai Arabica green coffee beans," *Smart Agricultural Technology*, vol. 9, p. 100680, Dec. 2024. [Online]. Available: https://doi.org/10.1016/j.atech.2024.100680

[12] W. Pinheiro Claro Gomes, L. Gonçalves, C. Barboza da Silva, and W. R. Melchert, "Application of multispectral imaging combined with machine learning models to discriminate special and traditional green coffee," *Computers and Electronics in Agriculture*, vol. 198, p. 107097, Jul. 2022. [Online]. Available: https://doi.org/10.1016/j.compag.2022.107097

[13] M. N. Izza and G. P. Kusuma, "Image classification of Green Arabica Coffee using transformer-based architecture," *International Journal of Engineering Trends and Technology*, vol. 72, no. 6, pp. 304–314, Jun. 2024. [Online]. Available: https://doi.org/10.14445/22315381/IJETT-V72I6P128

[14] H. F. Alhasson and S. S. Alharbi, "Classification of saudi coffee beans using a mobile application leveraging squeeze vision transformer technology," *Neural Computing and Applications*, vol. 37, no. 14, pp. 8629–8649, Feb. 2025. [Online]. Available: https://doi.org/10.1007/s00521-025-11024-9

[15] Y. Jiao, Y. Zhao, A. Jia, T. Wang, J. Li, K. Xiang, H. Deng, M. He, R. Jiang, and Y. Zhang, "Swin-HSSAM: a green coffee bean grading method by swin transformer," *PLOS One*, vol. 20, no. 5, p. e0322198, May 2025. [Online]. Available: https: //doi.org/10.1371/JOURNAL.PONE.0322198

[16] J. H. L. Goh, E. Ang, S. Srinivasan, X. Lei, J. Loh, T. C. Quek, C. Xue, X. Xu, Y. Liu, C.-Y. Cheng, J. C. Rajapakse, and Y.-C. Tham, "Comparative analysis of vision transformers and conventional convolutional neural networks in detecting referable diabetic retinopathy," *Ophthalmology Science*, vol. 4, no. 6, p. 100552, Nov. 2024. [Online]. Available: https: //doi.org/10.1016/j.xops.2024.100552

[17] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, "Swin Transformer V2: scaling up capacity and resolution," *arXiv*, 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2111.09883

[18] S. Studer, T. B. Bui, C. Drescher, A. Hanuschkin, L. Winkler, S. Peters, and K.-R. Müller, "Towards CRISP-ML(Q): a machine learning process model with quality assurance methodology," *Machine Learning and Knowledge Extraction*, vol. 3, no. 2, pp. 392–413, Apr. 2021. [Online]. Available: https://doi.org/10.3390/make3020020

[19] A. Febriana, K. Muchtar, R. Dawood, and C.-Y. Lin, "USK-Coffee dataset: A multi-class green arabica coffee bean dataset for deep learning," in *2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*. IEEE, Jun. 2022, pp. 469–473. [Online]. Available: https://doi.org/10.1109/CyberneticsCom55287.2022.9865489

[20] Patricio Bolívar Betancourt Ludeña, "Lojano Arabica coffee," *Zenodo*, 2025. [Online]. Available: https://doi.org/10.34740/kaggle/dsv/13947455

[21] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Oct. 2019. [Online]. Available: http://dx.doi.org/10.1007/s11263-019-01228-7

[22] H. L. Gope and H. Fukai, "Peaberry and normal coffee bean classification using CNN, SVM, and KNN: their implementation in and the limitations of Raspberry Pi 3," *AIMS Agriculture and Food*, vol. 7, no. 1, pp. 149–167, 2022. [Online]. Available: https://doi.org/10.3934/agrfood.2022010