Ingeniería Investiga

Vol. 8, e1400, año 2026
ISSN – Online: 2708-3039
DOI: https://doi.org/10.47796/ing.v8i00.1400

Artículo original

Aplicación de redes neuronales artificiales para la detección binaria del síndrome del ojo

Application of artificial neural networks for binary detection of red eye syndrome

Marcelino Torres Villan[1,*] | Juan Pedro Santos Fernández[2]

Afiliación:

[1,2]Escuela de Ingeniería de Sistemas, Universidad Nacional de Trujillo, La libertad, Perú

Autor de correspondencia: E-mail: *mtorres@unitru.edu.pe

Información del artículo:

Recibido: 17/01/2026
Aceptado: 04/03/2026
Publicado: 13/03/2026

Resumen

El síndrome del ojo rojo es uno de los motivos más frecuentes de consulta en atención primaria, y su diagnóstico temprano resulta complejo debido a la similitud clínica entre diversas etiologías. En este estudio se desarrolló y evaluó un enfoque de detección binaria (“ojo rojo” vs. “normal”) mediante la comparación de arquitecturas basadas en redes neuronales convolucionales (CNN), modelos basados en Transformers y un modelo híbrido. Se empleó un conjunto de 2 298 imágenes reorganizadas en dos clases, entrenadas bajo condiciones homogéneas mediante transfer learning y el uso de hiperparámetros fijos. Los experimentos se ejecutaron en Python 3.10.0 utilizando PyTorch 2.7.1+cu118, torchvision 0.22.1+cu118, timm 1.0.17, scikit-learn 1.6.1, NumPy 1.26.4, Albumentations 2.0.8 y Matplotlib 3.8.2, sobre un sistema con GPU NVIDIA RTX 4060 (8 GB). Los resultados evidenciaron un alto desempeño en todos los modelos evaluados (F1 > 0,92, MCC > 0,90 y AUC ≥ 0,98). El modelo híbrido alcanzó el mejor rendimiento global (AUC = 0,996, MCC = 0,925, F1 = 0,924 y exactitud = 94,20 %). La prueba de McNemar indicó que no existen diferencias estadísticamente significativas entre el modelo híbrido y el mejor modelo individual (ResNet).

Palabras clave: hiperemia ocular; aprendizaje profundo; visión computacional; clasificación binaria.

Abstract

Red eye syndrome is one of the most frequent reasons for consultation in primary care, and its early diagnosis is challenging due to the clinical similarity among different etiologies. In this study, a binary detection approach (“red eye” vs. “normal”) was developed and evaluated by comparing convolutional neural network (CNN) architectures, Transformer-based models, and a hybrid model. A dataset of 2,298 images reorganized into two classes was used and trained under homogeneous conditions using transfer learning and fixed hyperparameters. The experiments were conducted in Python 3.10.0 using PyTorch 2.7.1+cu118, torchvision 0.22.1+cu118, timm 1.0.17, scikit-learn 1.6.1, NumPy 1.26.4, Albumentations 2.0.8, and Matplotlib 3.8.2, on hardware equipped with an NVIDIA RTX 4060 GPU (8 GB). The results showed high performance across all evaluated models (F1 > 0.92, MCC > 0.90, and AUC ≥ 0.98). The hybrid model achieved the best overall performance (AUC = 0.996, MCC = 0.925, F1 = 0.924, and accuracy = 94.20%). McNemar’s test indicated no statistically significant differences between the hybrid model and the best-performing individual model (ResNet).

Keywords: ocular hyperemia; deep learning; computer vision; binary classification.

1. Introducción

El síndrome del ojo rojo constituye uno de los motivos de consulta más frecuentes en los servicios de atención primaria y de emergencia a nivel mundial. Es importante precisar que el "ojo rojo" no representa una enfermedad única, sino un signo o síndrome clínico observable, caracterizado principalmente por hiperemia conjuntival (inyección conjuntival) y asociado a múltiples etiologías. Su presentación clínica, que incluye hiperemia conjuntival, lagrimeo, dolor ocular, fotofobia y secreción, comparte signos comunes entre diversas patologías, lo que dificulta un diagnóstico diferencial preciso en el primer contacto clínico. Estudios recientes, como el de Sargolzaeimoghaddam (2025), señalan que cerca del 6 % de las atenciones en medicina general y el 15 % en oftalmología corresponden a cuadros de ojo rojo, lo que refleja su alta prevalencia y la complejidad de su abordaje inicial. Esta dificultad resulta aún más evidente en contextos donde el acceso a oftalmólogos es limitado y los profesionales dependen casi exclusivamente de la evaluación visual subjetiva.

La literatura internacional muestra que la variabilidad diagnóstica y el retraso en la identificación de entidades como conjuntivitis, queratitis, uveítis o glaucoma agudo incrementan el riesgo de tratamientos inadecuados, el uso innecesario de antibióticos y la progresión hacia complicaciones visuales severas. Tamimi et al. (2023) reportaron que aproximadamente el 20 % de los pacientes atendidos por dolor ocular y molestias inespecíficas fueron clasificados como casos de ojo rojo, lo que evidencia la necesidad de herramientas complementarias que permitan reducir la subjetividad clínica. Asimismo, Dag et al. (2024) destacaron que, en servicios de emergencia, un porcentaje considerable de consultas por ojo rojo no correspondía a verdaderas urgencias oftalmológicas, lo que revela ineficiencias en los procesos de triaje y priorización.

En el contexto peruano, estas dificultades se acentúan debido a limitaciones estructurales del sistema de salud. Se estima que millones de personas presentan algún grado de discapacidad visual o condiciones oculares no diagnosticadas oportunamente, lo que genera una elevada demanda de atención especializada. En regiones como La Libertad, donde el Instituto Regional de Oftalmología concentra una importante carga asistencial, la mayoría de establecimientos de salud carece de equipos tecnológicos avanzados o de personal especializado para realizar diagnósticos oportunos. Este escenario conduce a tratamientos empíricos, derivaciones tardías y sobrecarga en los centros de referencia, lo que afecta la calidad y la oportunidad del servicio. Además, factores ambientales como la presencia de partículas contaminantes y la exposición prolongada a pantallas digitales han incrementado los casos de irritación y enrojecimiento ocular, agravando la demanda asistencial.

Este estudio se fundamenta en la necesidad de integrar modelos avanzados de inteligencia artificial en el ámbito sanitario, fortaleciendo el cuerpo teórico de la ingeniería de sistemas aplicada al diagnóstico clínico. El desarrollo de redes neuronales artificiales contribuye al avance científico en áreas como visión computacional, ingeniería de software y ciencias de la salud, permitiendo la creación de herramientas que reduzcan la subjetividad diagnóstica y aporten mayor rigurosidad en la interpretación de imágenes oftálmicas. Desde esta perspectiva, el estudio representa un aporte significativo al conocimiento al explorar arquitecturas modernas que pueden mejorar la detección del síndrome del ojo rojo y generar evidencia sobre el potencial del aprendizaje profundo en entornos clínicos.

A nivel práctico y social, la propuesta responde a una necesidad real en los servicios de atención primaria, donde el personal médico enfrenta dificultades para diferenciar las causas del ojo rojo en ausencia de herramientas tecnológicas de apoyo. La implementación de un sistema automático de diagnóstico puede optimizar la precisión, reducir errores y agilizar la toma de decisiones, especialmente en zonas rurales o en centros con limitada disponibilidad de oftalmólogos. Asimismo, al facilitar diagnósticos oportunos, se pueden prevenir complicaciones visuales, reducir tratamientos innecesarios y promover una mayor equidad en el acceso a la salud visual.

En este marco, el desarrollo de un sistema inteligente de detección automática del síndrome del ojo rojo basado en redes neuronales se presenta como una alternativa estratégica. Este sistema permitiría reconocer patrones visuales asociados a distintos tipos de ojo rojo, generar alertas automáticas y priorizar derivaciones, contribuyendo a la optimización de recursos y a una atención médica más oportuna. En este contexto, el problema no se limita únicamente a la falta de personal especializado, sino también a la carencia de soluciones informáticas inteligentes que apoyen la toma de decisiones en tiempo real dentro del sistema de salud de Trujillo. En consecuencia, el objetivo general del presente estudio es desarrollar un modelo de redes neuronales artificiales para la detección binaria del síndrome del ojo rojo en imágenes oftálmicas. Para ello, se plantean los siguientes objetivos específicos: (1) entrenar modelos binarios basados en arquitecturas CNN y Transformers para clasificar imágenes como "ojo rojo" o "normal"; (2) evaluar el impacto de las arquitecturas DenseNet, ResNet, Inception, ViT y DeiT sobre el rendimiento del modelo; y (3) implementar un prototipo web funcional que integre el modelo con mejor desempeño para la detección automática del ojo rojo.

2. Metodología

2.1. Diseño del estudio

La investigación adoptó un diseño experimental computacional de tipo comparativo, orientado a evaluar el rendimiento de distintas arquitecturas de redes neuronales convolucionales (CNN) y modelos basados en Transformer para la detección binaria del síndrome del ojo rojo a partir de imágenes oftálmicas. Asimismo, se implementó un modelo híbrido que combina las arquitecturas consideradas en el estudio.

Cada arquitectura se consideró como una unidad experimental independiente y fue entrenada bajo las mismas condiciones de preprocesamiento, partición de datos y evaluación, con el fin de garantizar la reproducibilidad de los experimentos y permitir comparaciones válidas entre modelos.

2.2. Conjunto de datos

Se empleó una base de datos pública seleccionada por su disponibilidad, calidad visual y presencia de etiquetas clínicas, correspondiente al conjunto "Image Dataset on Eye Diseases Classification (Uveitis, Conjunctivitis, Cataract, Eyelid) with Symptoms and SMOTE Validation" (Bitto, 2024), disponible en Mendeley Data. El conjunto está conformado por 2 298 imágenes en formato JPG, obtenidas en condiciones de iluminación natural y clínica.

El corpus incluyó las siguientes clases originales: cataratas (544 imágenes), conjuntivitis (357), párpados caídos (525), normal (649) y uveítis (223), totalizando 2 298 imágenes. Para los fines de este estudio, las clases fueron reorganizadas en dos categorías con el fin de establecer un enfoque de detección binaria: Normal; imágenes etiquetadas como normal y Ojo rojo; imágenes correspondientes a las clases no normales del repositorio (cataratas, conjuntivitis, párpados caídos y uveítis).

Para la partición del conjunto de datos (N = 2 298) se realizó una división en entrenamiento, validación y prueba en proporción 70 % / 15 % / 15 %, mediante muestreo estratificado por clase, con el objetivo de preservar la proporción de imágenes "ojo rojo" y "normal" en cada subconjunto. Esta partición se aplicó de forma idéntica para todas las arquitecturas evaluadas, garantizando condiciones experimentales comparables.

Adicionalmente, se aplicó aumento de datos (data augmentation) con un factor ×10 exclusivamente sobre el conjunto de entrenamiento, con el objetivo de incrementar la variabilidad de las muestras durante el proceso de aprendizaje, sin modificar los conjuntos de validación ni de prueba.

2.3. Arquitecturas evaluadas

Se seleccionaron cinco modelos representativos de aprendizaje profundo, correspondientes a arquitecturas CNN y Transformers, además de un modelo híbrido.

a. Modelos CNN

DenseNet

Arquitectura caracterizada por la propagación eficiente de información entre capas, lo que reduce el problema del gradiente y mejora el aprendizaje de características visuales sutiles. Xu et al. (2023) reportaron que DenseNet-121 alcanzó un AUC aproximado de 0,998, con sensibilidad de 97,7 % y especificidad de 98,2 % en la detección de queratitis, lo que evidencia el alto potencial de esta arquitectura en aplicaciones oftalmológicas.

ResNet

Considerada una arquitectura de referencia en tareas de visión por computadora, debido a la incorporación de conexiones residuales que facilitan el entrenamiento de redes profundas. Hasan et al. (2025) desarrollaron un sistema explicable para diagnóstico y estadificación de glaucoma basado en ResNet, alcanzando un AUC de 0.96 (IC 95 %: 0.95–0.98), lo que demuestra su capacidad para tareas de diagnóstico clínico asistido por inteligencia artificial.

Inception

Arquitectura diseñada para capturar características a múltiples escalas mediante módulos paralelos de convolución. Pan et al. (2023) reportaron un desempeño competitivo de Inception en clasificación binaria de glaucoma frente a otras CNN en un conjunto amplio de imágenes de fondo de ojo, lo que respalda su utilidad en tareas de cribado oftalmológico

b. Modelos Transformer

Vision Transformer (ViT)

Primer modelo que adaptó la arquitectura Transformer al procesamiento de imágenes, dividiendo la entrada en parches y aplicando mecanismos de autoatención para aprender representaciones globales. Hui et al. (2024) demostraron que los Vision Transformers pueden superar a las CNN en el análisis de fotografías de fondo de ojo, evidenciando la capacidad de la autoatención para capturar patrones globales relevantes.

Data-efficient Image Transformer (DeiT)

Propuesto para abordar una limitación importante de los Vision Transformers: la necesidad de grandes volúmenes de datos para su entrenamiento. Le et al. (2024) reportaron que DeiT mantiene un desempeño competitivo incluso con conjuntos de datos de tamaño moderado, lo que facilita su aplicación en contextos clínicos donde los datasets suelen ser limitados.

c. Modelo híbrido CNN–Transformer

Con el propósito de combinar la capacidad de las CNN para extraer características locales con la habilidad de los Transformers para modelar dependencias globales, se evaluó un modelo híbrido CNN–Transformer. Esta aproximación busca integrar información espacial detallada con representaciones globales de la imagen. En este contexto, Rajatha y Ashoka (2025) propusieron el modelo híbrido EffiViT, el cual alcanzó un AUC de 0.9466 y un F1-score de 0.75, mostrando mejoras respecto a métodos previos. Estos resultados evidencian el potencial de las arquitecturas híbridas para tareas de clasificación binaria en imágenes médicas, lo que justifica su inclusión en el presente estudio para la detección automática del síndrome del ojo rojo.

2.4. Configuración de entrenamiento e Hiperparametros

La Tabla 1 presenta las arquitecturas específicas empleadas en el estudio para cada enfoque evaluado, incluyendo modelos basados en redes neuronales convolucionales (CNN) —ResNet-18, Inception-Next-Small y DenseNet-121— y modelos basados en Transformer —DeiT-Tiny y ViT-Tiny—. Estas arquitecturas fueron seleccionadas debido a su uso frecuente y al buen desempeño reportado en tareas de clasificación de imágenes médicas.

Tabla 1

Enfoque evaluado de cada arquitectura

Arquitectura	Tipo de modelo	Implementación
ResNet	CNN	resnet18
Inception	CNN	inception_next_small
DenseNet	CNN	densenet121
DeiT	Transformer	deit_tiny_patch16_224
ViT	Transformer	vit_tiny_patch16_224

Por otro lado, la Tabla 2 resume los hiperparámetros de entrenamiento aplicados de manera uniforme en todos los modelos, con el fin de garantizar condiciones experimentales homogéneas y permitir comparaciones válidas entre arquitecturas. Entre los parámetros considerados se incluyen el número máximo de épocas de entrenamiento, el tamaño de lote, la tasa de aprendizaje, el criterio de early stopping, el tamaño de entrada de la imagen y el factor de aumento de datos.

Tabla 2

Hiperparámetros de entrenamiento para cada modelo

Parámetro	Valor
Número máximo de épocas	30
Tamaño de lote (batch size)	32
Tasa de aprendizaje	$1{,}00 \times 10^{-4}$
Paciencia (early stopping)	3
Tamaño de entrada de imagen	224 × 224
Factor de aumento de datos	×10

Modelo híbrido: mecanismo de fusión

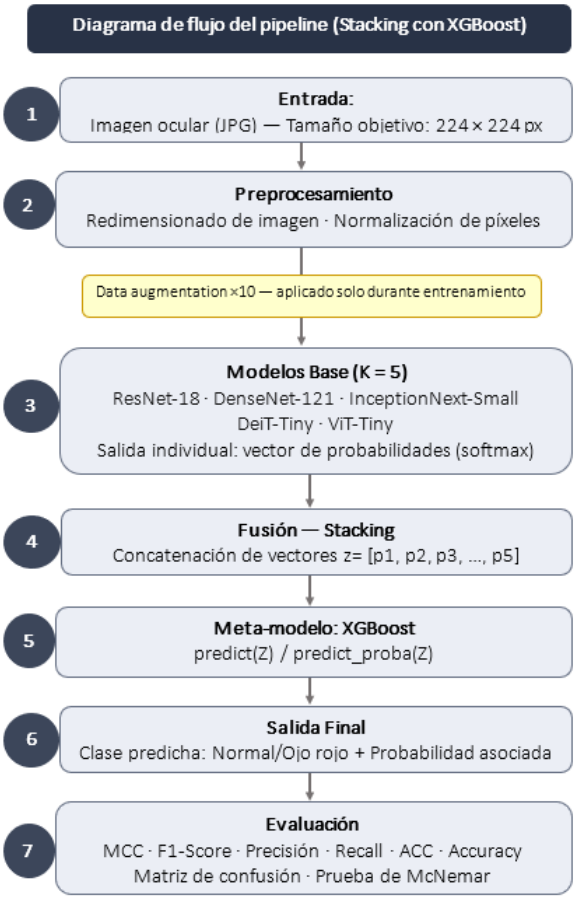
El modelo híbrido se implementó mediante un esquema de ensamble tipo stacking. En una primera etapa, cada arquitectura base (CNN y Transformers) genera una salida probabilística a través de la función softmax, produciendo un vector de probabilidades asociado a cada clase. Posteriormente, estos vectores se concatenan para formar un único vector de características, que se utiliza como entrada para un meta-clasificador basado en XGBoost, encargado de producir la predicción final.

El meta-modelo genera tanto la clase estimada como la probabilidad asociada a cada categoría, lo que permite integrar la información proveniente de todos los modelos base en una única decisión final más robusta y consistente.

La Figura 1 resume el pipeline completo: preprocesamiento, inferencia con modelos base (CNN y Transformers), concatenación de probabilidades (softmax) y fusión mediante stacking con XGBoost. El aumento de datos se aplica únicamente en el conjunto de entrenamiento.

Figura 1

Diagrama de flujo del pipeline y del modelo híbrido (stacking con XGBoost)



2.5. Métricas de evaluación

La evaluación del rendimiento se realizó mediante las siguientes métricas ampliamente utilizadas en visión médica:

Exactitud (Accuracy): Representa la proporción de predicciones correctas respecto al total de casos evaluados. Sin embargo, en presencia de desbalance de clases esta métrica puede resultar poco representativa; por ello, se complementa con métricas más robustas como MCC, F1-score y AUC. En estudios de imágenes médicas, Müller (2022) señala que la exactitud suele utilizarse como medida general de reconocimiento de patrones. No obstante, valores superiores al 90 % deben interpretarse con cautela cuando se analizan de forma aislada, ya que pueden ocultar problemas asociados al desbalance de clases.

$$Exactitud = \frac{VP+VN}{VP+VN+FP+FN} \quad (1)$$

Precisión (Precision): Se define como la proporción de casos correctamente identificados como positivos respecto al total de muestras clasificadas como positivas por el modelo. Devikala et al. (2025) destacan que esta métrica permite evaluar la confiabilidad de las predicciones positivas generadas por el clasificador.

$$Precisión = \frac{VP}{VP+FP} \quad (2)$$

Sensibilidad (Recall): También denominada recall, representa la capacidad del modelo para identificar correctamente los casos positivos. En el contexto clínico, esta métrica resulta especialmente relevante para detectar correctamente a los pacientes con la condición evaluada. Rainio et al. (2024) indican que la sensibilidad, junto con la especificidad, proporciona una visión más completa del desempeño del modelo, especialmente cuando existe desbalance entre clases.

$$Sensibilidad = \frac{VP}{VP+FN} \quad (3)$$

F1-score: El F1-score corresponde a la media armónica entre la precisión y la sensibilidad, lo que permite evaluar el equilibrio entre ambas métricas. Molina Arias (2024) señala que esta medida es especialmente útil en pruebas diagnósticas, ya que integra información tanto del valor predictivo positivo como de la capacidad de detección del modelo.

$$\text{Puntuación F1} = 2 * \frac{Precisión * Sensibilidad}{Precisión + Sensibilidad} \quad (4)$$

Matthews Correlation Coefficient (MCC): El MCC se considera una medida más completa para evaluar clasificadores binarios, ya que tiene en cuenta verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos en un único indicador. Chicco y Jurman (2020) demostraron que, en comparación con métricas tradicionales, el MCC mantiene mayor estabilidad estadística y proporciona una evaluación más equilibrada cuando se analizan datos clínicos complejos.

$$MCC = \frac{(VP*VN)-(FP*FN)}{\sqrt{(VP+FP)*(VP+FN)*(VN+FP)*(VN+FN)}} \quad (5)$$

Área bajo la curva ROC (AUC): El área bajo la curva ROC evalúa la capacidad del modelo para discriminar entre clases positivas y negativas a diferentes umbrales de decisión. Reifs Jiménez et al. (2025) indican que esta métrica resulta especialmente útil en conjuntos de datos con desbalance entre clases.

$$AUC = \int_0^1 S(FP)dFP \quad (6)$$

Estas métricas permiten evaluar no solo el rendimiento global del modelo, sino también su estabilidad en entornos clínicos donde minimizar los falsos negativos resulta prioritario. Además, se generaron matrices de confusión y curvas ROC para visualizar el comportamiento de los clasificadores sobre el conjunto de prueba.

2.6. Análisis estadístico

Para determinar si existían diferencias estadísticamente significativas entre los modelos CNN, Transformer y el modelo híbrido, se aplicó la prueba de McNemar, la cual permite evaluar si las discrepancias observadas entre dos clasificadores se deben al azar o reflejan diferencias reales en su capacidad de clasificación.

Para la implementación de esta prueba se emplearon las siguientes librerías de Python:

- *Statsmodels:* Utilizada para la ejecución formal de la prueba de McNemar y el cálculo del estadístico χ^2 y su valor p, proporcionando una base estadística robusta.
- *Scipy:* Empleada como respaldo para funciones estadísticas complementarias y validación de resultados.
- *Numpy:* Utilizada para la manipulación eficiente de arreglos y la construcción de las tablas de contingencia a partir de las predicciones de los modelos.
- *Pandas*: Empleada para la organización y gestión estructurada de los resultados experimentales y métricas comparativas.

2.7. Ambiente computacional

Los experimentos se ejecutaron en un entorno computacional compuesto por un procesador Intel Core i9-13900H con 14 núcleos y 20 hilos a una frecuencia base de 2,60 GHz, acompañado de una GPU NVIDIA RTX 4060 con 8 GB de memoria. El sistema utilizó una unidad de almacenamiento SSD NVMe Micron 3400 de 512 GB con velocidades aproximadas de lectura y escritura de 6600/3600 MB/s. Además, se contó con 32 GB de memoria RAM DDR5 (2 × 16 GB) a 5200 MT/s, lo que permitió ejecutar eficientemente los procesos de entrenamiento y evaluación de los modelos.

El desarrollo y entrenamiento de los modelos se realizó en Visual Studio Code versión 1.102.1, utilizando Python versión 3.10.0. Para el entrenamiento de redes neuronales se empleó PyTorch (torch 2.7.1+cu118) junto con torchvision 0.22.1+cu118 para tareas de visión por computadora. Asimismo, se utilizaron las siguientes bibliotecas: timm 1.0.17, NumPy 1.26.4 y pandas 2.1.3 para procesamiento de datos; scikit-learn 1.6.1 para métricas y utilidades de evaluación; albumentations 2.0.8 para aumento de datos; matplotlib 3.8.2 para visualización; tqdm 4.67.1 para monitoreo del progreso de entrenamiento; y psutil 7.0.0 y GPUtil 1.4.0 para el monitoreo de recursos del sistema.

Para la generación de reportes se utilizaron Pillow 10.4.0 para el manejo de imágenes y reportlab 4.4.0 para la creación de documentos PDF. El módulo datetime (incluido en Python) se empleó para la gestión de fechas y horas. Finalmente, la aplicación web desarrollada para el despliegue del sistema se implementó utilizando HTML, CSS y JavaScript.

3. Resultados

Se evaluaron cinco arquitecturas de aprendizaje profundo: tres redes convolucionales (DenseNet, ResNet e Inception) y dos modelos basados en Transformer (ViT y DeiT), todas ajustadas para la clasificación binaria del síndrome del ojo rojo. Para cada modelo se seleccionó la mejor época de entrenamiento en función del Matthews Correlation Coefficient (MCC), debido a su mayor robustez frente a posibles desbalances de clase.

En términos globales, como se observa en la Tabla 3, todos los modelos alcanzaron un rendimiento elevado, con valores de F1 superiores a 0,92, AUC cercanos o superiores a 0,98 y MCC por encima de 0,90, lo que evidencia una alta capacidad de discriminación entre las clases “ojo rojo” y “normal.

Estos resultados evidencian un desempeño alto y consistente entre las arquitecturas evaluadas. La Figura 2 presenta la comparación gráfica de las principales métricas, donde se observa que todos los modelos superan umbrales comúnmente aceptados en clasificación médica (F1 > 0,90 y MCC > 0,85), lo que indica una capacidad estable para discriminar entre imágenes con ojo rojo y ojos normales.

El análisis de la evolución durante el entrenamiento muestra que el modelo DeiT, seleccionado como referencia para ilustrar el comportamiento del aprendizaje, redujo su pérdida de entrenamiento de 0,74 a 0,29 entre las épocas 1 y 6, mientras que el F1-score aumentó de 0,89 a 0,94 y el MCC de 0,86 a 0,92. A partir de este punto, las mejoras adicionales fueron marginales y se observaron ligeras oscilaciones en las métricas, lo que sugiere el inicio de un posible sobreajuste. Este comportamiento indica que un número moderado de épocas de entrenamiento (entre 6 y 10) resulta suficiente para alcanzar un rendimiento estable, pudiéndose aplicar estrategias de early stopping para reducir el tiempo de cómputo sin afectar significativamente el desempeño.

Las matrices de confusión asociadas a la mejor época de cada modelo evidenciaron un elevado número de aciertos en ambas clases y un número reducido de falsos negativos, aspecto especialmente relevante en contextos clínicos donde omitir un caso positivo puede tener consecuencias importantes para el paciente. En particular, ResNet y DeiT mostraron una proporción equilibrada de verdaderos positivos y verdaderos negativos, consistente con sus valores elevados de MCC.

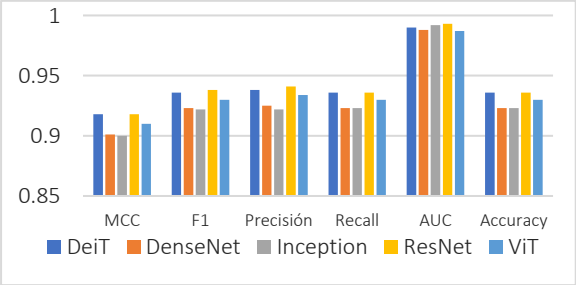
Tabla 3

Resumen de resultados

Modelos	Época	Pérdida Entrenamiento	MCC Actual	F1	Precisión	Recall	AUC	Accuracy
DeiT	6	0,291	0,918	0,936	0,938	0,936	0,990	0,936
DenseNet	10	0,340	0,901	0,923	0,925	0,923	0,988	0,923
Inception	7	0,312	0,900	0,922	0,922	0,923	0,992	0,923
ResNet	9	0,330	0,918	0,938	0,941	0,936	0,993	0,936
Vit	9	0,352	0,910	0,930	0,934	0,930	0,987	0,930

Figura 2

Resultados de los modelos según las métricas



3.1. Rendimiento computacional

Además del desempeño predictivo, se evaluó el rendimiento computacional de los modelos durante el proceso de entrenamiento. La Tabla 4 presenta métricas relacionadas con el tiempo de entrenamiento, uso de CPU, consumo de memoria RAM y utilización de GPU para la mejor época de cada arquitectura.

Como se observa en la Figura 3, los tiempos de entrenamiento por época fueron comparables entre arquitecturas, con ligeras variaciones. Inception presentó el mayor tiempo de entrenamiento, mientras que ResNet fue el modelo más rápido. Por su parte, los modelos Transformer (ViT y DeiT) se ubicaron en un rango intermedio debido a su mayor complejidad computacional. No obstante, el rendimiento obtenido por DeiT en términos de métricas de clasificación compensó parcialmente este costo adicional. En contraste, ResNet mostró un equilibrio favorable entre eficiencia computacional y desempeño predictivo, lo que coincide con su estabilidad reportada en diversas aplicaciones de clasificación médica

3.2. Resultados del modelo híbrido

El modelo híbrido, basado en la combinación de arquitecturas CNN (ResNet, Inception y DenseNet) y modelos Transformer (DeiT y ViT), mostró uno de los mejores desempeños globales entre los enfoques evaluados. Como se presenta en la Tabla 5, el modelo alcanzó valores elevados y equilibrados en las métricas de evaluación, destacando por su alta capacidad discriminativa y estabilidad en la clasificación binaria del síndrome del ojo rojo.

Asimismo, estos resultados indican que la integración de información local (CNN) y global (Transformers) contribuye a mejorar el desempeño general frente a los modelos individuales. La Figura 4 presenta una comparación global de las métricas obtenidas por cada modelo, evidenciando la ventaja del enfoque híbrido en términos de capacidad discriminativa.

Tabla 4

Rendimiento computacional de los modelos

Modelo	Época óptima	Tiempo entrenamiento (s)	Tiempo validación (s)	CPU entrenamiento (%)	RAM entrenamiento (%)	GPU entrenamiento (%)	VRAM entrenamiento (MB)
DeiT	6	138,64	44,08	7,08	55,43	40,35	385
DenseNet	10	138,45	44,75	13,71	51,41	34,96	645
Inception	7	157,41	46	6,79	51,02	46,11	939
ResNet	9	132,31	54,04	23,41	61,77	15,92	1 390,0
ViT	9	137,8	44,14	7,17	55,01	33,88	385

Figura 3

Gráfico de rendimiento

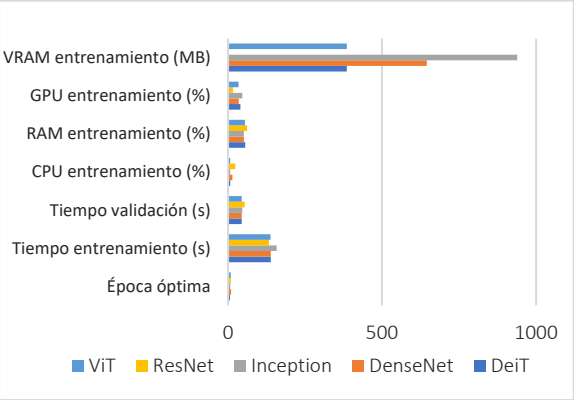


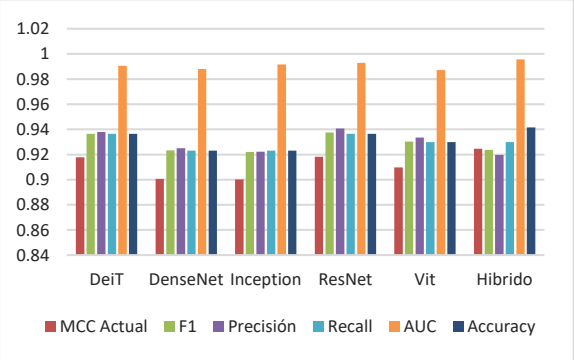
Tabla 5

Resultado del modelo hibrido

MCC Actual	F1	Precisión	Recall	AUC	Accuracy
0,925	0,924	0,919	0,929	0,996	0,942

Figura 4

Resumen de las métricas de cada modelo en general de todos



3.3. Comparación estadística entre modelos

Con el fin de fortalecer la evaluación, se aplicó la prueba de McNemar, cuyos resultados se presentan en la Tabla 6. El análisis indicó que no existen diferencias estadísticamente significativas entre el modelo híbrido y ResNet (p = 0,1489).

Tabla 6

Resultados de la prueba de McNemar para la comparación entre ResNet y el modelo híbrido

χ^2	p-value	Interpretación
2,0833	0,1489	No hay diferencia significativa en los modelos

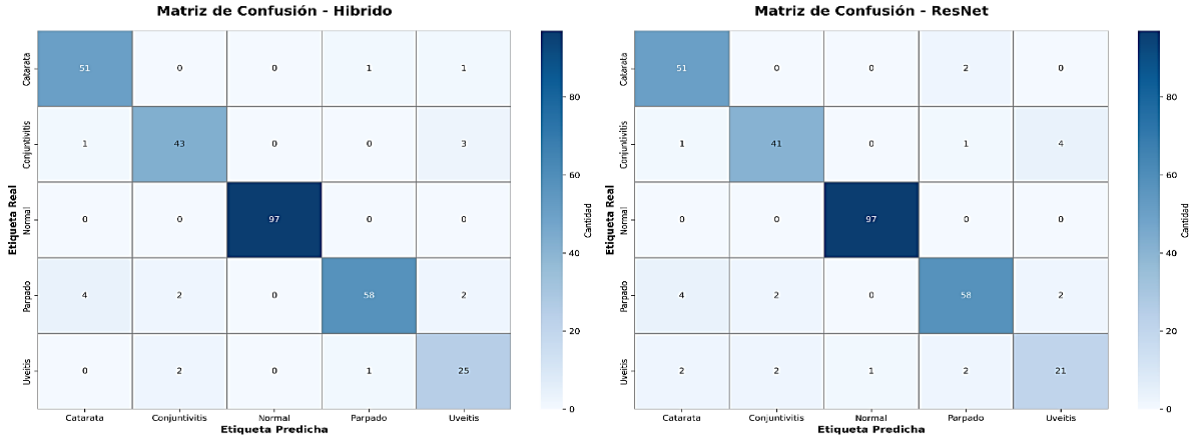
Este resultado sugiere que ambos modelos presentan un comportamiento de clasificación comparable sobre el mismo conjunto de prueba, proporcionando respaldo estadístico a la comparación entre arquitecturas CNN, Transformers y el enfoque híbrido.

3.4. Matrices de confusión y prototipo del sistema

La Figura 5 presenta las matrices de confusión correspondientes a los dos modelos con mejor desempeño global: el modelo híbrido y ResNet. En ambos casos se observa una adecuada concentración de predicciones en la diagonal principal, lo que indica una alta concordancia entre las etiquetas reales y las predichas. El modelo híbrido muestra una ligera reducción en los errores de clasificación, evidenciando una mejor integración de la información generada por los modelos base.

Figura 5

Matrices de confusión de los mejores modelos (Hibrido – Resnet)



Finalmente, la Figura 6 muestra la interfaz del prototipo web desarrollado para cargar imágenes y visualizar la predicción generada por el sistema (clase estimada y probabilidad asociada). Esta figura se incluye como evidencia de implementación del sistema; no obstante, la validación del desempeño del modelo se sustenta principalmente en las métricas cuantitativas, matrices de confusión y pruebas estadísticas presentadas.

Figura 6
Interfaz del sistema Eyeres

Como complemento, el sistema genera un reporte automático en formato PDF que resume la predicción final del modelo híbrido junto con la imagen evaluada. Este reporte permite una rápida revisión del resultado y facilita la trazabilidad del análisis, aunque la evaluación científica del desempeño se fundamenta en las métricas y análisis estadísticos previamente descritos

4. Discusión

Los resultados obtenidos evidencian un desempeño alto y consistente en las cinco arquitecturas evaluadas, con valores de MCC ≥ 0,900, F1 ≥ 0,922 y AUC ≥ 0,987. En particular, el modelo híbrido alcanzó un rendimiento equilibrado y competitivo, destacando por su elevada capacidad discriminativa (AUC) y estabilidad global (MCC). Estos resultados respaldan la hipótesis de que la combinación de arquitecturas CNN y Transformers puede aportar mayor robustez en tareas de clasificación clínica, especialmente cuando las señales visuales son sutiles y heterogéneas, como ocurre en imágenes del segmento anterior del ojo.

Al contrastar estos hallazgos con la literatura científica, se observa coherencia con estudios que reportan altos niveles de desempeño en el análisis de imágenes oftálmicas. Por ejemplo, Li et al. (2021) reportaron un rendimiento cercano al máximo teórico en la detección de queratitis, con AUC de 0,998, sensibilidad de 97,7 % y especificidad de 98,2 %. Aunque el objetivo clínico y el conjunto de datos utilizados en dicho estudio difieren del presente trabajo, esta comparación sugiere que, cuando se dispone de datos bien anotados y representativos, los modelos de aprendizaje profundo pueden alcanzar niveles elevados de discriminación diagnóstica.

Desde una perspectiva más amplia, el metaanálisis realizado por Ong et al. (2024) sobre queratitis infecciosa estimó una sensibilidad de 86,2 % y especificidad de 96,3 % en evaluaciones externas frente al estándar clínico de referencia, señalando además una comparabilidad significativa con el desempeño de oftalmólogos en ciertos subanálisis. Estos resultados son relevantes para la interpretación de los hallazgos del presente estudio, ya que evidencian que el rendimiento observado en evaluaciones internas suele disminuir cuando los modelos se enfrentan a conjuntos de datos externos. En este sentido, aunque los resultados obtenidos muestran un desempeño sólido en el conjunto evaluado, un paso necesario para fortalecer la validez clínica del sistema sería realizar validaciones externas o esquemas de validación cruzada estratificados por fuente o condiciones de captura.

De manera similar, Ueno et al. (2024) desarrollaron un sistema de inteligencia artificial para la detección de múltiples enfermedades del segmento anterior utilizando imágenes capturadas con teléfonos inteligentes. En su estudio reportaron AUC de 0,986 para queratitis infecciosa y AUC de 0,992 para cataratas, así como valores cercanos a 1,0 en otras categorías diagnósticas. Aunque dicho trabajo aborda un

problema multiclase y emplea un pipeline diferente, sus resultados respaldan la capacidad de los enfoques basados en aprendizaje profundo para mantener un alto nivel de discriminación incluso en escenarios de adquisición de imágenes más variables, como los obtenidos mediante dispositivos móviles. Este aspecto resulta especialmente relevante para la propuesta del presente estudio, orientada hacia el desarrollo de sistemas de apoyo clínico potencialmente desplegables.

En relación con la comparación entre arquitecturas CNN, Vision Transformers y modelos híbridos, el estudio de Zhang et al. (2025) sobre retinopatía diabética (tarea multiclase) reportó que los enfoques híbridos pueden alcanzar un desempeño más equilibrado al integrar características locales y globales. En su trabajo, el mejor modelo obtuvo una exactitud de 72,93 % y un coeficiente QWK de 0,841, reflejando la complejidad de la clasificación por niveles de severidad. Aunque este contexto difiere del problema abordado en el presente estudio (clasificación binaria en imágenes del segmento anterior), dichos resultados aportan evidencia adicional de que las arquitecturas híbridas pueden mejorar la robustez y estabilidad de los modelos al combinar distintos mecanismos de representación visual.

A pesar de los resultados prometedores, este estudio presenta algunas limitaciones que deben ser consideradas. En primer lugar, la recategorización binaria utilizada se definió de manera operativa a partir de las etiquetas disponibles en el repositorio, agrupando distintos diagnósticos en la clase “ojo rojo”. Por lo tanto, la clase positiva no representa una etiología única y puede incluir variabilidad clínica entre patologías. En segundo lugar, el conjunto de datos proviene de una fuente pública y podría no reflejar completamente la variabilidad presente en escenarios clínicos reales, incluyendo diferencias en dispositivos de captura, condiciones de iluminación o niveles de severidad de la enfermedad. En tercer lugar, no se realizó una validación externa multicéntrica, por lo que el rendimiento observado podría variar al aplicarse en contextos clínicos distintos. Finalmente, el uso de un modelo híbrido basado en ensamble incrementa el costo computacional del sistema, lo que podría limitar su implementación en entornos con recursos tecnológicos restringidos.

Como líneas de trabajo futuro, se plantea ampliar el conjunto de datos incorporando imágenes provenientes de diferentes fuentes clínicas, realizar validaciones externas multicéntricas y explorar técnicas de calibración de probabilidades y métodos de explicabilidad (XAI) que permitan interpretar las decisiones del modelo. Estas estrategias contribuirían a fortalecer la confiabilidad del sistema y facilitar su potencial integración como herramienta de apoyo en la práctica clínica.

5. Conclusiones

El Los resultados obtenidos demuestran que la inteligencia artificial, aplicada mediante redes neuronales profundas, constituye una herramienta eficaz para la detección binaria del síndrome del ojo rojo a partir de imágenes oftálmicas. Todos los modelos evaluados alcanzaron un alto desempeño, con valores de AUC superiores a 0,98 y MCC mayores a 0,90, lo que confirma la capacidad de los enfoques de aprendizaje profundo para analizar de manera confiable este tipo de patologías.

El modelo híbrido, basado en la combinación de arquitecturas CNN y Transformers, obtuvo el mejor rendimiento global, alcanzando un AUC de 0,996, un MCC de 0,925, un F1-score de 0,924 y una exactitud de 94,20 %. Estos resultados evidencian que la integración de características locales y globales contribuye a mejorar la estabilidad y la capacidad discriminativa del sistema frente a modelos individuales.

El análisis estadístico mediante la prueba de McNemar indicó que no se observaron diferencias estadísticamente significativas entre el modelo híbrido y el modelo individual de mejor desempeño (ResNet), lo que sugiere un comportamiento consistente entre ambas aproximaciones en la clasificación de imágenes oftálmicas.

Desde una perspectiva clínica, los resultados respaldan el potencial de la inteligencia artificial como herramienta de apoyo al diagnóstico médico, al facilitar la detección

temprana del síndrome del ojo rojo. En este sentido, el sistema propuesto podría contribuir a mejorar los procesos de cribado, telemedicina y apoyo a la toma de decisiones clínicas, especialmente en contextos donde el acceso a especialistas en oftalmología es limitado.

Contribución de los autores

M. Torres: Conceptualización, curación de datos, análisis formal, investigación, metodología, Administración del proyecto, recursos, software, supervisión, visualización y redacción del borrador original. J. P. Santos: Conceptualización, análisis formal, adquisición de fondos, administración del proyecto, supervisión, validación, redacción del borrador original y revisión y edición del manuscrito.

Conflictos de interés

Los autores declaran no tener ningún conflicto de interés relacionado con esta publicación.

6. Referencias bibliográficas

Bitto, A. K. (2024). Image Dataset on Eye Diseases Classification (Uveitis, Conjunctivitis, Cataract, Eyelid) with Symptoms and SMOTE Validation. *Mendeley Data*, *2*. https://doi.org/10.17632/n9zp473wfw.2

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21(1). https://doi.org/10.1186/s12864-019-6413-7

Dag, Y., Seyfi Aydın, & Ebrar Kumantas. (2024). The profile of patients attending to the general emergency department with ocular complaints within the last year: is it a true ocular emergency? *BMC Ophthalmology*, *24*(1). https://doi.org/10.1186/s12886-024-03608-1

Devikala, S., Vinoth, S., Shaby, S. M., Govindaraju, A. B., Vidhya, K., & Vijayalakshmi, K. (2025). A Multi-Component Attention Graph Convolutional Neural Network Optimized by the Gooseneck Barnacle Algorithm for High-Precision ECG Arrhythmia Classification in Sensor-Based Biomedical Systems. *Biomedical Signal Processing and Control*, *113*, 108866. https://doi.org/10.1016/j.bspc.2025.108866

Hasan, M. M., Phu, J., Wang, H., Sowmya, A., Kalloniatis, M., & Meijering, E. (2025). OCT-based diagnosis of glaucoma and glaucoma stages using explainable machine learning. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-87219-w

Hui, J., Ang, E., Srinivasan, S., Lei, X., Loh, J., Quek, T. C., Xue, C., Xu, X., Liu, Y., Cheng, C.-Y., Rajapakse, J. C., & Tham, Y.-C. (2024). Comparative Analysis of Vision Transformers and Conventional Convolutional Neural Networks in Detecting Referable Diabetic Retinopathy. *Ophthalmology Science*, *4*(6), 100552–100552. https://doi.org/10.1016/j.xops.2024.100552

Le, N. T., Le Truong, T., Deelertpaiboon, S., Srisiri, W., Pongsachareonnont, P. F., Suwajanakorn, D., Mavichak, A., Itthipanichpong, R., Asdornwised, W., Benjapolakul, W., Chaitusaney, S., & Kaewplung, P. (2024). ViT-AMD: A New Deep Learning Model for Age-Related Macular Degeneration Diagnosis From Fundus Images. *International Journal of Intelligent Systems*, *2024*(1). https://doi.org/10.1155/2024/3026500

Li, Z., Jiang, J., Chen, K., Chen, Q., Zheng, Q., Liu, X., Weng, H., Wu, S., & Chen, W. (2021). Preventing corneal blindness caused by keratitis using artificial intelligence. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-24116-6

Molina Arias, M. (2024). Un intruso de otro mundo: F1-score. Revista Electrónica AnestesiaR, 16(4), 3. https://doi.org/10.30445//rear.v16i4.1258

Müller, D., Soto-Rey, I., & Kramer, F. (2022). Towards a guideline for evaluation metrics in medical image segmentation. *BMC Research Notes*, *15*(1). https://doi.org/10.1186/s13104-022-06096-y

Ong, Z. Z., Sadek, Y., Qureshi, R., Liu, S.-H., Li, T., Liu, X., Takwoingi, Y., Sounderajah, V., Ashrafian, H., Ting, D. S. W., Mehta, J. S., Rauz, S., Said, D. G., Dua, H. S., Burton, M. J., & Ting, D. S. J. (2024). Diagnostic performance of deep learning for infectious keratitis: a systematic review and meta-analysis. *EClinicalMedicine*, *77*, 102887. https://doi.org/10.1016/j.eclinm.2024.102887

Pan, Y., Liu, J., Cai, Y., Yang, X., Zhang, Z., Long, H., Zhao, K., Yu, X., Zeng, C., Duan, J., Xiao, P., Li, J., Cai, F., Yang, X., & Tan, Z. (2023). Fundus image classification using Inception V3 and ResNet-50 for the early diagnostics of fundus diseases. *Frontiers in Physiology*, *14*. https://doi.org/10.3389/fphys.2023.1126780

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-56706-x

Rajatha, & Ashoka, D. V. (2025). EffiViT: Hybrid CNN-Transformer for Retinal Imaging. *Computers in Biology and Medicine*, *191*, 110164. https://doi.org/10.1016/j.compbiomed.2025.110164

Reifs Jiménez, D., Casanova-Lozano, L., Grau-Carrión, S., & Reig-Bolaño, R. (2025). Artificial Intelligence Methods for Diagnostic and Decision-Making Assistance in Chronic Wounds: A Systematic Review. *Journal of Medical Systems*, *49*(1). https://doi.org/10.1007/s10916-025-02153-8

Sargolzaeimoghaddam, M., Maral Sargolzaeimoghaddam, Kothari, Z., Sebhat, A. M., & Soleimani, M. (2025). Review of ophthalmic emergencies in primary care: a comprehensive approach to red eye. Annals of Eye Science, 10, 20–20. https://doi.org/10.21037/aes-25-10

Tamimi, A., Allawi, M. N., & Kishore Hanumantharayappa. (2023). Characterization of red eye cases presented to the eye emergency clinic at a tertiary care hospital during COVID-19 Pandemic. *Oman Journal of Ophthalmology*, *16*(2), 220–226. https://doi.org/10.4103/ojo.ojo_224_22

Ueno, Y., Oda, M., Yamaguchi, T., Fukuoka, H., Nejima, R., Kitaguchi, Y., Miyake, M., Akiyama, M., Miyata, K., Kashiwagi, K., Maeda, N., Shimazaki, J., Noma, H., Mori, K., & Oshika, T. (2024). Deep learning model for extensive smartphone-based diagnosis and triage of cataracts and multiple corneal diseases. *British Journal of Ophthalmology*, *108*(10), 1406–1413. https://doi.org/10.1136/bjo-2023-324488

Xu, Z., Xu, J., Shi, C., Xu, W., Jin, X., Han, W., Jin, K., Grzybowski, A., & Yao, K. (2023). Artificial Intelligence for Anterior Segment Diseases: A Review of Potential Developments and Clinical Applications. *Ophthalmology and Therapy*, *12*(3), 1439–1455. https://doi.org/10.1007/s40123-023-00690-4

Zhang, W., Belcheva, V., & Ermakova, T. (2025). Interpretable Deep Learning for Diabetic Retinopathy: A Comparative Study of CNN, ViT, and Hybrid Architectures. *Computers*, *14*(5), 187–187. https://doi.org/10.3390/computers14050187