



Detección temprana de glaucoma mediante modelos híbridos basados en aprendizaje profundo en imágenes de fondo de ojo

Early glaucoma detection using hybrid models based on deep learning in fundus images

Ramírez-Cenepo, Leandro Enrique^{1*}

Ruiz-Cueva, John Antony¹

Sánchez-Flores, Jhosep¹

Rojas-Córdova, Kelvin Lleins¹

Leveau-Paredes, Monica Gabriela¹

Llontop-Reategui, Augusto Ricardo¹

¹Grupo de Investigación en Inteligencia Artificial, Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional de San Martín, Tarapoto, Perú

²Grupo de Investigación en Salud, Educación, Responsabilidad Social, Facultad de Medicina Humana, Universidad Nacional de San Martín, Tarapoto, Perú

Recibido: 12 Ene. 2026 | **Aceptado:** 22 Jun. 2026 | **Publicado:** 20 Jul. 2026

Autor de correspondencia*: lramirezc@alumno.unsm.edu.pe

Cómo citar este artículo: Ramírez-Cenepo, L. E., Ruiz-Cueva, J. A., Sánchez-Flores, J., Rojas-Córdova, K. J., Leveau-Paredes & M.G., Llontop-Reategui, A.R. (2026). Detección temprana de glaucoma mediante modelos híbridos basados en aprendizaje profundo en imágenes de fondo de ojo. *Revista Científica de Sistemas e Informática*, 6(2), e1642. <https://doi.org/10.51252/rcsi.v6i2.1642>

RESUMEN

El glaucoma es una de las principales causas de ceguera irreversible, por lo que su detección temprana es esencial para prevenir la pérdida visual. La investigación tuvo como objetivo desarrollar y evaluar modelos híbridos de aprendizaje profundo para la clasificación multiclase del glaucoma mediante imágenes de fondo de ojo. Se utilizaron 6014 imágenes retinianas de bases públicas, sometidas a depuración, balanceo y partición estratificada. Las arquitecturas VGG16, ResNet50 y EfficientNet-B0 se emplearon como extractores de características combinados con Cascaded Deep Forest. La evaluación incluyó validación cruzada estratificada de 10 pliegues y una prueba independiente. VGG16 + CDF obtuvo el mejor desempeño, con una accuracy promedio de 0.9320, mayor estabilidad y mejor generalización. Las pruebas estadísticas confirmaron diferencias significativas entre los modelos, demostrando que el enfoque híbrido constituye una alternativa robusta para la clasificación automatizada del glaucoma.

Palabras clave: análisis de imágenes retinianas; clasificación multiclase; diagnóstico automatizado; transferencia de aprendizaje; visión por computadora

ABSTRACT

Glaucoma is one of the leading causes of irreversible blindness, making early detection essential to prevent vision loss. This study aimed to develop and evaluate hybrid deep learning models for multiclass glaucoma classification using fundus images. A total of 6,014 retinal images from public databases were used, after undergoing cleaning, balancing, and stratified splitting. The VGG16, ResNet50, and EfficientNet-B0 architectures were employed as feature extractors combined with Cascaded Deep Forest. Evaluation included stratified 10-fold cross-validation and an independent test set. VGG16 + CDF achieved the best performance, with an average accuracy of 0.9320, greater stability, and better generalization. Statistical tests confirmed significant differences among the models, demonstrating that the proposed hybrid approach is a robust alternative for automated glaucoma classification.

Keywords: retinal image analysis; multiclass classification; automated diagnosis; transfer learning; computer vision



1. INTRODUCCIÓN

El glaucoma constituye una de las principales causas de ceguera irreversible a nivel mundial y representa un problema prioritario de salud pública debido a su progresión crónica y silenciosa. Esta neuropatía óptica se caracteriza por el deterioro progresivo del nervio óptico, generalmente asociado a un aumento de la presión intraocular, lo que conduce a una pérdida irreversible del campo visual periférico. Se estima que la carga global de la enfermedad continúa en aumento, impulsada por el envejecimiento poblacional y las limitaciones en la detección temprana (Schuster et al., 2020; Tham et al., 2014). La naturaleza asintomática del glaucoma en sus etapas iniciales provoca que un alto porcentaje de casos sea diagnosticado en fases avanzadas, cuando el daño visual ya es irreversible.

La detección oportuna del glaucoma resulta fundamental para prevenir la pérdida permanente de la visión; sin embargo, los métodos diagnósticos convencionales presentan importantes limitaciones. Técnicas como la tonometría, la tomografía de coherencia óptica y la perimetría automatizada requieren equipamiento especializado y personal altamente capacitado, lo que restringe su accesibilidad en contextos con recursos limitados (Ha & Park, 2019; Realini et al., 2021; Sidhu & Mansoori, 2024). Asimismo, la interpretación de imágenes oftalmológicas depende en gran medida de la experiencia del especialista, lo que introduce variabilidad diagnóstica y aumenta el riesgo de errores clínicos (Chaurasia et al., 2025).

En los últimos años, la inteligencia artificial y, en particular, el aprendizaje profundo ha emergido como herramientas prometedoras para el desarrollo de sistemas automatizados de apoyo al diagnóstico médico. Diversos estudios evidencian que las redes neuronales convolucionales permiten identificar patrones complejos en imágenes de fondo de ojo con altos niveles de precisión, alcanzando desempeños comparables a los de especialistas humanos en tareas específicas (Chiang et al., 2024; Hwang et al., 2025; Ling et al., 2025). Además, el uso de técnicas de transferencia de aprendizaje ha facilitado la implementación de modelos robustos incluso en escenarios con disponibilidad limitada de datos etiquetados (Govindharaj et al., 2025). De manera complementaria, aplicaciones recientes basadas en aprendizaje automático han evidenciado el potencial de estas técnicas para la identificación de patrones complejos y la construcción de modelos predictivos en distintos contextos científicos (Arévalo-Hernández et al., 2023).

No obstante, a pesar de estos avances, gran parte de los enfoques propuestos se basa en datos unimodales, principalmente imágenes de fondo de ojo, lo que limita la representación integral de la complejidad estructural y funcional del glaucoma (Shi et al., 2025). Adicionalmente, persisten desafíos relacionados con la calidad heterogénea de las imágenes, la escasez de conjuntos de datos clínicos bien anotados y la presencia de sesgos en los datos de entrenamiento, lo que afecta la capacidad de generalización de los modelos (Ilesanmi et al., 2023; Świerczyński et al., 2025). Estas limitaciones reducen la confiabilidad de los sistemas propuestos en escenarios clínicos reales.

La problemática se intensifica en países de ingresos bajos y medianos, donde la disponibilidad de especialistas en oftalmología es limitada y el acceso a servicios diagnósticos es insuficiente (Meethal et al., 2024). En el contexto peruano, el glaucoma afecta aproximadamente al 2 % de la población mayor de 40 años, con una alta proporción de casos no diagnosticados oportunamente, especialmente en zonas rurales y urbano-marginales (Ministerio de Salud, 2021). Esta situación evidencia la necesidad de soluciones tecnológicas accesibles que permitan ampliar la cobertura diagnóstica y fortalecer las estrategias de prevención de la ceguera evitable.

A pesar del creciente número de investigaciones en el área, se identifica un vacío científico relevante: la falta de modelos de aprendizaje profundo robustos, comparativamente evaluados y validados en condiciones reales que alcancen niveles de precisión diagnóstica clínicamente útiles en poblaciones locales. Asimismo, existe limitada evidencia sobre el desempeño relativo de diferentes arquitecturas de redes neuronales convolucionales aplicadas a la detección temprana del glaucoma utilizando imágenes de fondo de ojo (Sharma et al., 2025; Shoukat et al., 2023).

En este contexto, el objetivo de la presente investigación fue desarrollar y evaluar modelos híbridos basados en aprendizaje profundo para la detección temprana de glaucoma a partir de imágenes de fondo de ojo, con el propósito de identificar la configuración con mejor capacidad de generalización en la clasificación multiclase de la enfermedad.

2. MATERIALES Y MÉTODOS

La presente investigación corresponde a un estudio de tipo aplicado, de nivel explicativo, con diseño no experimental y de corte transversal. Su propósito fue desarrollar y evaluar modelos híbridos basados en aprendizaje profundo para la detección temprana de glaucoma a partir de imágenes de fondo de ojo. El enfoque metodológico integró redes neuronales convolucionales preentrenadas como extractores de características y un clasificador basado en *Cascaded Deep Forest*, con el fin de analizar su capacidad para discriminar diferentes estadios de la enfermedad (Aljohani & Aburasain, 2024).

El desarrollo experimental se implementó en el lenguaje de programación Python, empleando *PyTorch* y *timm* para la extracción de características mediante redes neuronales convolucionales preentrenadas; *scikit-learn* para la implementación de métricas de evaluación y procedimientos de validación; y *Optuna* para la optimización de hiperparámetros. Asimismo, se utilizaron *NumPy*, *Pandas*, *Matplotlib* y *Seaborn* para el procesamiento, análisis y visualización de los resultados. Los experimentos se ejecutaron en Google Colab con soporte de GPU para acelerar el procesamiento computacional.

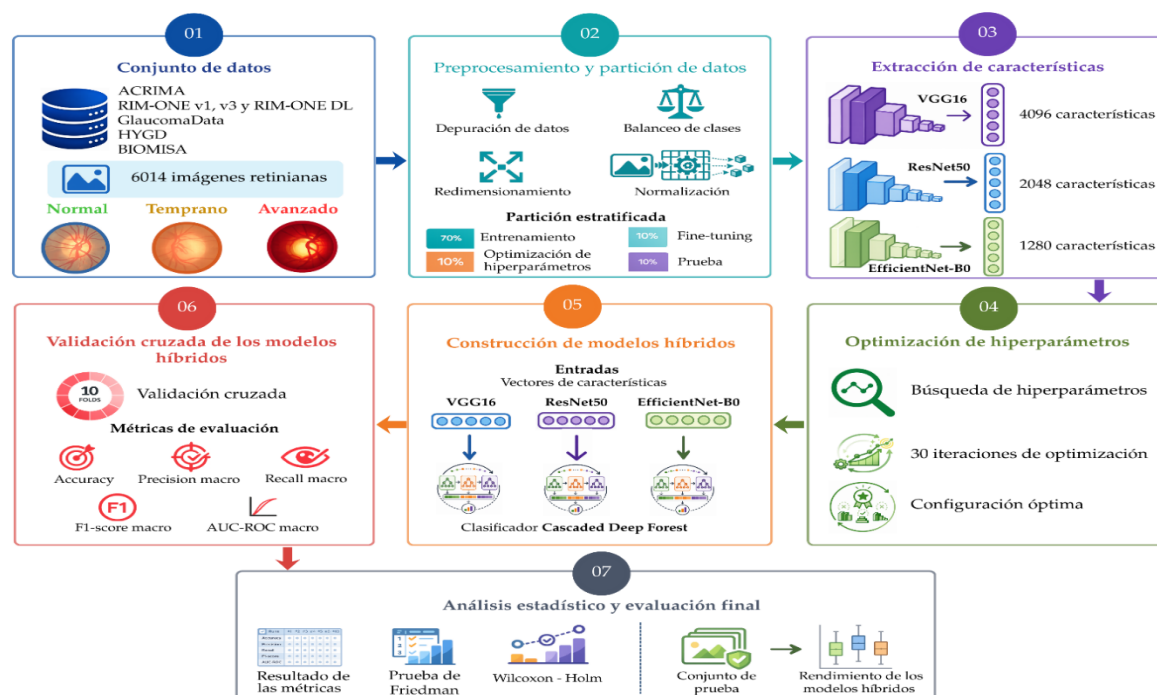


Figura 1. Flujo metodológico del estudio

El proceso metodológico desarrollado en la investigación integró las etapas de preparación y procesamiento de datos, construcción y evaluación de modelos híbridos basados en aprendizaje profundo, así como procedimientos de validación y análisis estadístico orientados a evaluar su desempeño en la clasificación multiclase del glaucoma (**Figura 1**).

Fase 1: Conjunto de datos

El estudio utilizó imágenes de fondo de ojo obtenidas mediante la integración de múltiples conjuntos de datos públicos especializados en glaucoma, incluyendo ACRIMA (Díaz-Pinto et al., 2019), RIM-ONE v1, v3 y RIM-ONE DL (Fumero Batista et al., 2020), GlaucomaData disponible en Roboflow Universe (Roboflow, 2026), HYGD (Abramovich et al., 2026) y BIOMISA (Hassan et al., 2018). Estos conjuntos contienen imágenes retinianas clasificadas en diferentes estadios del glaucoma y fueron desarrollados para tareas de clasificación automatizada mediante técnicas de inteligencia artificial.

El conjunto de datos integrado estuvo conformado por 6014 imágenes distribuidas en las categorías normal (3223 imágenes), glaucoma temprano (1120 imágenes) y glaucoma avanzado (1671 imágenes). La **Figura 2** presenta ejemplos representativos de las categorías diagnósticas utilizadas en la investigación.

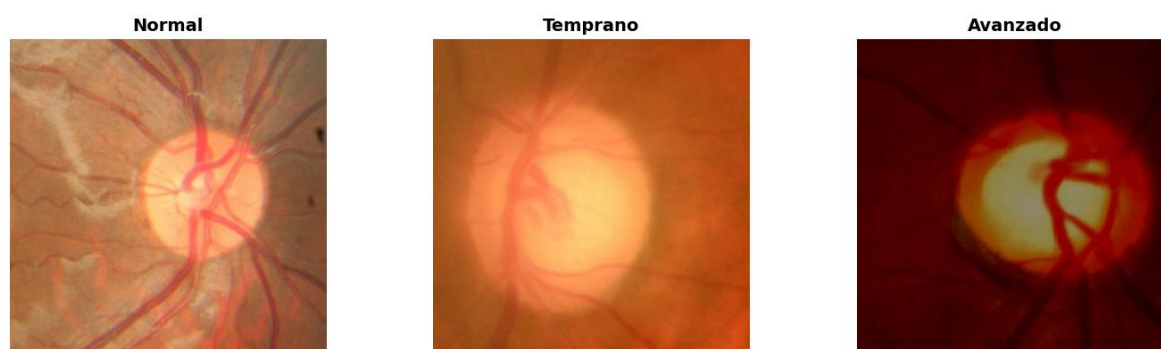


Figura 2. Muestras representativas del conjunto de datos

Fase 2: Preprocesamiento y partición de datos

Las imágenes fueron sometidas a un proceso de preprocesamiento orientado a garantizar la consistencia del conjunto de datos y la compatibilidad con las arquitecturas de aprendizaje profundo utilizadas. Inicialmente, se realizó una etapa de depuración que incluyó la verificación de etiquetas, la eliminación de registros duplicados y la revisión de imágenes inconsistentes o de baja calidad. Como resultado de este proceso, el conjunto de datos se redujo de 6014 a 5478 imágenes válidas para el desarrollo experimental (Saha et al., 2023).

Posteriormente, con el propósito de reducir el sesgo asociado al desbalance entre categorías diagnósticas, se aplicó un proceso de balanceo mediante selección aleatoria, conformando un conjunto equilibrado de 820 imágenes por clase y un total de 2460 imágenes balanceadas (Owusu-Adjei et al., 2023).

Las imágenes seleccionadas se redimensionaron a 224×224 píxeles, se normalizaron utilizando los parámetros de ImageNet y se convirtieron a tensores para su procesamiento computacional. Este procedimiento permitió estandarizar las entradas y garantizar compatibilidad con las arquitecturas convolucionales preentrenadas empleadas en la investigación.

Finalmente, el conjunto balanceado se dividió de manera estratificada en cuatro subconjuntos independientes: entrenamiento (70 %), fine-tuning (10 %), optimización de hiperparámetros (10 %) y prueba (10 %). Esta estrategia permitió ajustar los modelos, optimizar su configuración y preservar un conjunto de prueba completamente independiente para la evaluación final de la capacidad de generalización.

Fase 3: Extracción de características

Se utilizaron las arquitecturas VGG16, ResNet50 y EfficientNet-B0 como extractores de características profundas mediante transferencia de aprendizaje. Inicialmente, los modelos preentrenados en ImageNet fueron adaptados utilizando el subconjunto destinado al fine-tuning, permitiendo ajustar parcialmente sus representaciones al dominio de imágenes retinianas utilizadas en la investigación. Posteriormente, se eliminaron las capas de clasificación originales y se conservaron únicamente las representaciones generadas por las capas convolucionales internas (He et al., 2016; Simonyan & Zisserman, 2014; Tan & Le, 2019).

Este enfoque permitió aprovechar patrones visuales previamente aprendidos en grandes conjuntos de datos y, al mismo tiempo, especializar las representaciones profundas para la clasificación multiclase del glaucoma, reduciendo además la complejidad computacional asociada al entrenamiento completo de redes neuronales profundas.

Las representaciones profundas generadas por cada arquitectura fueron transformadas en vectores de características utilizados como entrada para el clasificador híbrido. Las dimensiones de los vectores variaron según la arquitectura empleada: VGG16 generó vectores de 4096 características, ResNet50 de 2048 y EfficientNet-B0 de 1280.

Fase 4: Optimización de hiperparámetros

La configuración de los modelos híbridos se optimizó mediante el algoritmo de búsqueda bayesiana implementado en Optuna. Durante esta etapa, se exploraron múltiples combinaciones de hiperparámetros asociados al clasificador Cascaded Deep Forest, incluyendo el número de capas en cascada, número de estimadores, profundidad máxima de los árboles, número mínimo de muestras por hoja y estrategias de selección de características (Akiba et al., 2019).

La optimización se realizó utilizando el subconjunto destinado al ajuste de hiperparámetros, con el propósito de identificar las configuraciones con mejor desempeño para cada arquitectura convolucional evaluada. Como criterio principal de optimización se empleó la accuracy, priorizando la capacidad global de clasificación en el problema multiclase del glaucoma.

El proceso de búsqueda se desarrolló mediante 30 iteraciones de optimización, permitiendo explorar diferentes configuraciones del espacio de hiperparámetros y seleccionar aquellas con mejor desempeño para cada arquitectura híbrida evaluada. El conjunto de prueba se mantuvo completamente independiente durante esta etapa, evitando filtración de información y reduciendo sesgos en la evaluación final de la capacidad de generalización de los modelos.

Fase 5: Construcción de modelos híbridos

Una vez identificadas las configuraciones óptimas durante la etapa de optimización, se construyeron los modelos híbridos integrando las arquitecturas convolucionales previamente descritas para la extracción de características profundas y un clasificador basado en Cascaded

Deep Forest (CDF) para la etapa de clasificación. Las representaciones vectoriales generadas por los extractores de características fueron utilizadas como entrada de los modelos híbridos.

El clasificador Cascaded Deep Forest está compuesto por múltiples capas secuenciales de modelos tipo Random Forest y Extra Trees, organizadas bajo una estructura en cascada. En cada nivel, las probabilidades de clase generadas por los clasificadores se concatenaron con los vectores de características originales, permitiendo construir representaciones progresivamente más discriminativas (Breiman, 2001; Geurts et al., 2006; Zhou & Feng, 2019).

Este enfoque permitió combinar la capacidad de extracción automática de características del aprendizaje profundo con la capacidad de generalización de los métodos basados en árboles de decisión, favoreciendo un aprendizaje más robusto para la clasificación multiclase del glaucoma.

Fase 6: Validación cruzada de los modelos híbridos

Una vez definidas las configuraciones óptimas durante la etapa de optimización, los modelos híbridos fueron evaluados mediante validación cruzada estratificada de 10 pliegues (10-fold), aplicada sobre el conjunto de entrenamiento. Este procedimiento permitió preservar la proporción de las categorías diagnósticas en cada partición y obtener una evaluación más robusta y consistente del desempeño de los modelos.

En cada pliegue se calcularon métricas orientadas a evaluar la capacidad de clasificación multiclase, incluyendo accuracy, precision macro, recall macro, F1-score macro y AUC-ROC macro. Posteriormente, los resultados obtenidos en las diferentes iteraciones fueron promediados para estimar el comportamiento general de cada modelo híbrido evaluado durante el proceso de validación (Hicks et al., 2022).

Fase 7: Análisis estadístico y evaluación final

A partir de las métricas obtenidas durante la validación cruzada estratificada de 10 pliegues, se realizó un análisis estadístico inferencial utilizando los valores de accuracy como criterio principal de comparación entre los modelos híbridos VGG16 + CDF, ResNet50 + CDF y EfficientNet-B0 + CDF. Para la comparación global entre configuraciones se empleó la prueba no paramétrica de Friedman, debido a que las mismas particiones fueron utilizadas en la evaluación de los modelos. Posteriormente, se aplicaron comparaciones pareadas mediante la prueba de Wilcoxon con corrección de Holm, con el fin de identificar diferencias específicas entre arquitecturas (Demšar, 2006).

De manera complementaria, cada modelo fue evaluado sobre el conjunto de prueba independiente, el cual no fue utilizado durante el entrenamiento, la optimización ni la validación cruzada. Esta evaluación permitió estimar la capacidad final de generalización de los modelos en un escenario completamente independiente.

3. RESULTADOS Y DISCUSIÓN

3.1. Distribución final del conjunto de datos

El proceso de depuración y balanceo permitió obtener un conjunto de datos más consistente y equilibrado para el desarrollo experimental. La **Tabla 1** presenta la distribución de imágenes durante las diferentes etapas del preprocesamiento aplicado.

Tabla 1. Distribución del conjunto de datos durante el proceso de depuración y balanceo

Etapa	Normal	Glaucoma temprano	Glaucoma avanzado	Total
Conjunto de datos integrado	3223	1120	1671	6014
Conjunto de datos depurado	2979	828	1671	5478
Conjunto de datos balanceado	820	820	820	2460

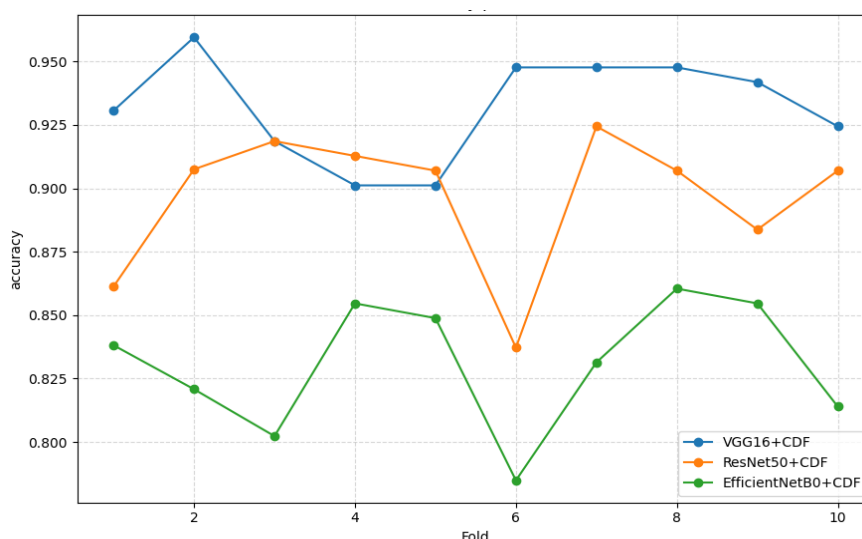
3.2. Desempeño de los modelos

La validación cruzada estratificada evidenció diferencias en el desempeño de los modelos híbridos evaluados para la clasificación multiclase del glaucoma. La **Tabla 2** presenta los resultados promedio obtenidos para las métricas de accuracy, precision macro, recall macro, F1-score macro y AUC-ROC macro. En términos de accuracy, VGG16 + CDF alcanzó el mayor valor promedio (0.9320), seguido por ResNet50 + CDF (0.8966) y EfficientNetB0 + CDF (0.8310). Una tendencia similar se observó en las demás métricas de evaluación, donde VGG16 + CDF presentó consistentemente los valores más altos durante la validación cruzada. Estas diferencias evidencian variaciones en la capacidad de las arquitecturas para representar patrones discriminativos asociados a las categorías diagnósticas del glaucoma.

Tabla 2. Resultados promedio en la validación cruzada estratificada de 10 pliegues

Modelo	Accuracy	Precision macro	Recall macro	F1-score macro	AUC-ROC macro
VGG16 + CDF	0.932041	0.932804	0.932002	0.932036	0.991054
ResNet50 + CDF	0.896646	0.903249	0.896572	0.896878	0.980329
EfficientNetB0 + CDF	0.831012	0.835778	0.831075	0.831260	0.948292

La variación de la accuracy entre los diferentes pliegues mostró diferencias en la estabilidad de los modelos híbridos evaluados (**Figura 3**). En términos generales, VGG16 + CDF presentó el comportamiento más consistente a lo largo de las iteraciones, manteniendo los valores más altos de accuracy en la mayoría de los pliegues analizados. Por su parte, ResNet50 + CDF mostró una variabilidad intermedia, mientras que EfficientNetB0 + CDF registró las mayores fluctuaciones y los menores valores relativos durante el proceso experimental.

**Figura 3.** Comportamiento de la accuracy durante la validación cruzada estratificada

Las matrices de confusión evidenciaron patrones diferenciados de clasificación entre las arquitecturas híbridas evaluadas (**Figura 4**). VGG16 + CDF concentró la mayor cantidad de aciertos sobre la diagonal principal, destacando especialmente en las categorías glaucoma avanzado y

glaucoma temprano, con 541 y 519 clasificaciones correctas, respectivamente. ResNet50 + CDF mantuvo un desempeño competitivo, aunque presentó un mayor número de confusiones entre las categorías glaucoma temprano y normal. Por su parte, EfficientNetB0 + CDF registró la mayor dispersión de errores, particularmente en la identificación de casos de glaucoma temprano, donde se observaron confusiones más frecuentes con las demás categorías diagnósticas.

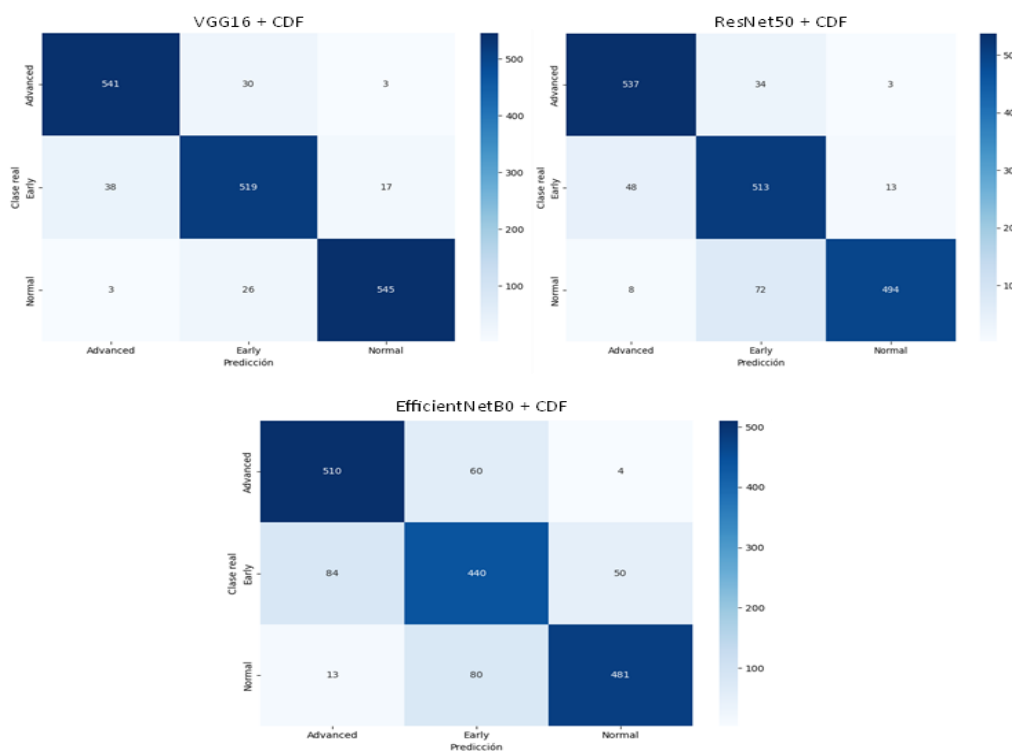


Figura 4. Matrices de confusión por modelo

3.3. Comparación estadística entre modelos

Con el propósito de determinar si las diferencias observadas entre los modelos híbridos evaluados eran estadísticamente significativas, se realizó un análisis inferencial utilizando los valores de accuracy obtenidos durante la validación cruzada estratificada de 10 pliegues. **Tabla 3** presenta los estadísticos descriptivos correspondientes a cada arquitectura evaluada. En términos generales, VGG16 + CDF registró la media de accuracy más alta y el intervalo de confianza más elevado entre los modelos analizados, mientras que ResNet50 + CDF mostró un comportamiento intermedio. Por su parte, EfficientNetB0 + CDF presentó los menores valores promedio durante el proceso de validación cruzada.

Tabla 3. Estadísticos descriptivos del accuracy

Modelo	Media	Desv. estándar	IC 95% inferior	IC 95% superior
VGG16 + CDF	0.9320	0.0204	0.9175	0.9466
ResNet50 + CDF	0.8966	0.0277	0.8768	0.9165
EfficientNetB0 + CDF	0.8310	0.0252	0.8130	0.8490

Previamente al análisis comparativo, se evaluó la distribución de los valores de accuracy mediante la prueba de Shapiro-Wilk. Los resultados evidenciaron que al menos una de las arquitecturas evaluadas no cumplía el supuesto de normalidad ($p < 0.05$), por lo que se optó por emplear pruebas no paramétricas para la comparación estadística entre modelos.

La prueba de Friedman evidenció diferencias estadísticamente significativas entre los modelos híbridos evaluados (**Tabla 4**). Los resultados obtenidos mostraron un valor de $\chi^2 = 16.67$ y un nivel de significancia de $p = 0.00024$, indicando diferencias relevantes en el desempeño de las arquitecturas durante la validación cruzada estratificada. Asimismo, el coeficiente de concordancia de Kendall ($W = 0.83$) evidenció un alto nivel de consistencia en el ordenamiento de los modelos a través de los diferentes pliegues evaluados.

Tabla 4. Resultado de la prueba de Friedman

Métrica	χ^2 de Friedman	p-valor	Kendall's W
accuracy	16.666667	0.00024	0.833333

Las comparaciones post hoc mediante la prueba de Wilcoxon con corrección de Holm evidenciaron diferencias estadísticamente significativas entre las arquitecturas híbridas evaluadas (**Tabla 5**). En particular, VGG16 + CDF presentó diferencias significativas respecto a ResNet50 + CDF y EfficientNetB0 + CDF, registrando además la mayor diferencia media frente a esta última arquitectura. Asimismo, ResNet50 + CDF también mostró diferencias significativas en comparación con EfficientNetB0 + CDF. Estos resultados confirman las diferencias observadas previamente durante la validación cruzada y respaldan la superioridad relativa de VGG16 + CDF en términos de desempeño global.

Tabla 5. Comparaciones post hoc Wilcoxon con corrección de Holm

Comparación	Diferencia media	p crudo	p ajustado (Holm)
VGG16 + CDF vs ResNet50 + CDF	0.035395	0.019531	0.019531
VGG16 + CDF vs EfficientNetB0 + CDF	0.101028	0.001953	0.005859
ResNet50 + CDF vs EfficientNetB0 + CDF	0.065634	0.001953	0.005859

La distribución de los valores de accuracy obtenidos durante la validación cruzada evidenció diferencias consistentes entre las arquitecturas híbridas evaluadas (**Figura 5**). VGG16 + CDF presentó la mediana más alta y una menor dispersión relativa en comparación con los demás modelos, mientras que EfficientNetB0 + CDF mostró los valores más bajos y una mayor variabilidad entre pliegues. Estos resultados refuerzan las diferencias identificadas mediante las pruebas estadísticas inferenciales.

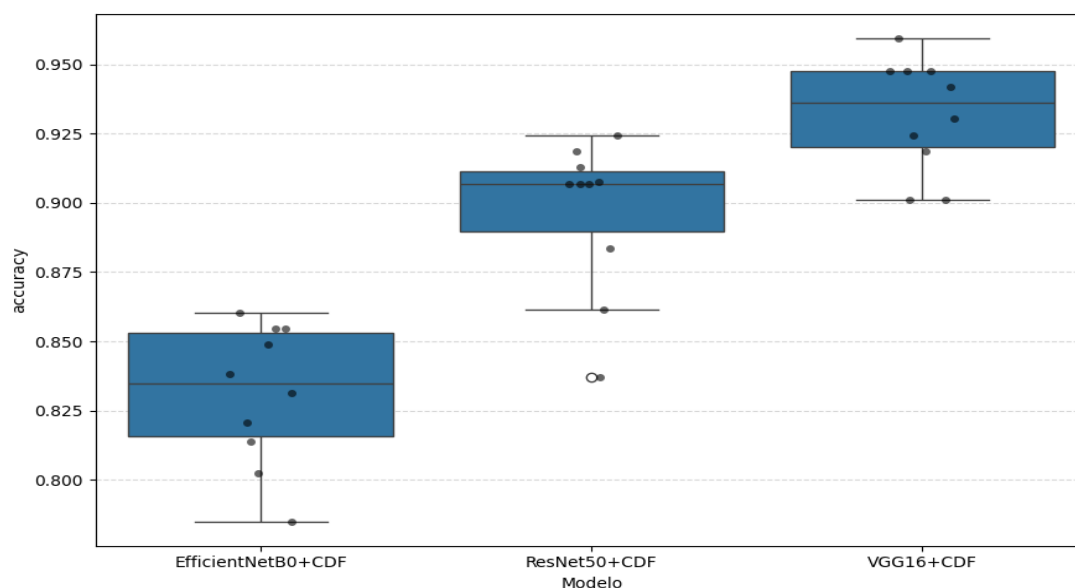


Figura 5. Distribución del accuracy por modelo

3.4. Generalización en el conjunto de prueba

La evaluación sobre el conjunto de prueba independiente evidenció diferencias en la capacidad de generalización de los modelos híbridos evaluados (**Figura 6**). En términos generales, VGG16 + CDF mantuvo los valores de accuracy más altos tanto durante la validación cruzada como en el conjunto de prueba, mostrando además una menor dispersión relativa entre pliegues. ResNet50 + CDF presentó un comportamiento intermedio, mientras que EfficientNetB0 + CDF registró el menor rendimiento global y una mayor variabilidad relativa. Asimismo, se observó una disminución de la accuracy en el conjunto de prueba respecto a los valores promedio obtenidos durante la validación cruzada, comportamiento esperado al evaluar imágenes no utilizadas durante el entrenamiento y la optimización de los modelos. Estos resultados evidencian que VGG16 + CDF presentó la capacidad de generalización más consistente para la clasificación multiclase del glaucoma.

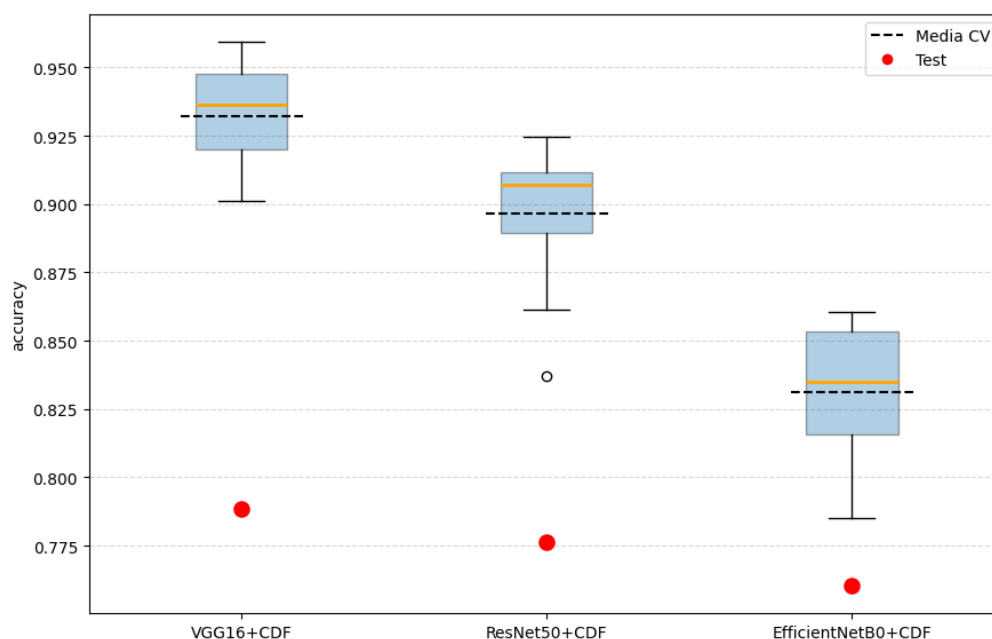


Figura 6. Rendimiento de los modelos en el conjunto de prueba

3.5. Discusión

Los modelos híbridos evaluados demostraron un desempeño competitivo para la clasificación multiclase del glaucoma a partir de imágenes de fondo de ojo, evidenciando que la combinación entre extracción de características profundas y clasificación mediante Cascaded Deep Forest constituye una estrategia efectiva para la detección automatizada de la enfermedad. Entre las arquitecturas analizadas, VGG16 + CDF presentó el mejor desempeño tanto en validación cruzada como en el conjunto de prueba independiente, mostrando además una mayor estabilidad entre pliegues y una capacidad de generalización más consistente respecto a ResNet50 + CDF y EfficientNetB0 + CDF.

Los resultados obtenidos guardan relación con lo reportado por Sharma et al. (2025), quienes desarrollaron un sistema híbrido para el cribado automatizado de glaucoma mediante imágenes de fondo de ojo, alcanzando elevados niveles de sensibilidad y especificidad. De manera similar, la presente investigación evidenció que la integración de arquitecturas convolucionales con Cascaded Deep Forest permitió mantener un desempeño robusto durante la validación cruzada y la evaluación sobre datos no vistos. Asimismo, Chaurasia et al. (2025) señalaron que el

rendimiento de los modelos tiende a disminuir al evaluarse con conjuntos externos o independientes, comportamiento que también fue observado en este estudio al comparar los resultados de validación cruzada con el conjunto de prueba.

En comparación con enfoques basados exclusivamente en redes convolucionales profundas, los resultados sugieren ventajas importantes del modelo híbrido propuesto. Shoukat et al. (2023) reportaron altos niveles de accuracy utilizando ResNet50 para la detección de glaucoma; sin embargo, en esta investigación la combinación ResNet50 + CDF presentó un rendimiento inferior respecto a VGG16 + CDF. Este comportamiento evidencia que el desempeño de una arquitectura convolucional puede variar según la estrategia de clasificación utilizada y las características del conjunto de datos empleado. Además, las matrices de confusión mostraron una mayor dificultad para discriminar entre las categorías glaucoma temprano y glaucoma avanzado, situación coherente con la complejidad clínica de la enfermedad y con la presencia de patrones visuales parcialmente solapados entre estadios.

Aunque investigaciones recientes como las de Hwang et al. (2025) y Świerczyński et al. (2025) destacan el potencial de los enfoques multimodales y multisensoriales para fortalecer el diagnóstico automatizado de glaucoma, los resultados alcanzados en esta investigación evidencian que es posible obtener un desempeño competitivo utilizando únicamente imágenes de fondo de ojo y modelos híbridos de menor complejidad computacional. Esto incrementa el potencial de aplicabilidad del enfoque propuesto en entornos con recursos limitados y plantea como línea futura la incorporación de estrategias multimodales para fortalecer aún más la capacidad de generalización clínica del sistema.

CONCLUSIONES

El enfoque híbrido, basado en redes neuronales convolucionales para extraer características y Cascaded Deep Forest para la clasificación, demostró un desempeño competitivo en la clasificación multiclase del glaucoma mediante imágenes de fondo de ojo. Los resultados evidenciaron que la combinación de aprendizaje profundo y métodos basados en árboles es efectiva para su detección automatizada. Entre las arquitecturas evaluadas, VGG16 + CDF presentó el mejor desempeño en la validación cruzada y el conjunto de prueba independiente, con mayor estabilidad y capacidad de generalización que ResNet50 + CDF y EfficientNetB0 + CDF. Asimismo, las pruebas estadísticas confirmaron diferencias significativas entre los modelos, respaldando la robustez del enfoque.

Las matrices de confusión mostraron mayor dificultad para diferenciar el glaucoma temprano del avanzado, debido a la complejidad de la progresión de la enfermedad. No obstante, el modelo presenta potencial como herramienta de apoyo para el tamizaje automatizado. Se recomienda incorporar enfoques multimodales, técnicas de interpretabilidad y conjuntos de datos más diversos para mejorar su generalización y aplicabilidad clínica.

FINANCIAMIENTO

Los autores no recibieron ningún patrocinio para llevar a cabo este estudio-artículo.

CONFLICTO DE INTERESES

No existe ningún tipo de conflicto de interés relacionado con la materia del trabajo.

CONTRIBUCIÓN DE AUTORES

Conceptualización: Ramírez-Cenepo, L. E.; Rojas-Córdova, K. L.; Llontop-Reategui, A. R. Curación de datos: Ramírez-Cenepo, L. E.; Ruiz-Cueva, J. A.; Sánchez-Flores, J. Análisis formal: Ramírez-Cenepo, L. E.; Rojas-Córdova, K. L. Investigación: Ramírez-Cenepo, L. E.; Ruiz-Cueva, J. A.; Sánchez-Flores, J.; Llontop-Reategui, A. R. Metodología: Ramírez-Cenepo, L. E.; Rojas-Córdova, K. L.; Leveau-Paredes, M. G. Administración del proyecto: Ramírez-Cenepo, L. E. Software: Sánchez-Flores, J.; Ramírez-Cenepo, L. E. Validación: Ramírez-Cenepo, L. E.; Sánchez-Flores, J. Visualización: Rojas-Córdova, K. L.; Leveau-Paredes, M. G. Redacción – borrador original: Ramírez-Cenepo, L. E.; Rojas-Córdova, K. L. Redacción – revisión y edición: Ramírez-Cenepo, L. E.; Ruiz-Cueva, J. A.; Leveau-Paredes, M. G.

REFERENCIAS

- Abramovich, O., Pizem, H., Fhima, J., Berkowitz, E., Gofrit, B., Baskin, M., Meisel, M., Eijgen, J. Van, Blumenthal, E., & Behar, J. (2026). Hillel Yaffe Glaucoma Dataset (HYGD): A Gold-Standard Annotated Fundus Dataset for Glaucoma Detection. *PhysioNet*.
<https://doi.org/https://doi.org/10.13026/m92s-0z95>
- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631.
<https://doi.org/10.1145/3292500.3330701>
- Aljohani, A., & Aburasain, R. Y. (2024). A hybrid framework for glaucoma detection through federated machine learning and deep learning models. *BMC Medical Informatics and Decision Making*, 24(1), 115. <https://doi.org/10.1186/s12911-024-02518-y>
- Arévalo-Hernández, C. O., Arévalo-Gardini, E., Arévalo-López, L. A., Tuesta-Hidalgo, O., Romero-Vela, D. S., & Ruiz-Camus, C. E. (2023). Predicción de la fertilidad del suelo mediante aprendizaje automático en la provincia de Alto Amazonas, Perú. *Revista Peruana de Investigación Agropecuaria*, 3(2), e63. <https://doi.org/10.56926/repia.v3i2.63>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
<https://doi.org/10.1023/A:1010933404324>
- Chaurasia, A. K., Liu, G.-S., Greatbatch, C. J., Gharahkhani, P., Craig, J. E., Mackey, D. A., MacGregor, S., & Hewitt, A. W. (2025). A generalised computer vision model for improved glaucoma screening using fundus images. *Eye*, 39(1), 109–117. <https://doi.org/10.1038/s41433-024-03388-4>
- Chiang, Y.-Y., Chen, C.-L., & Chen, Y.-H. (2024). Deep Learning Evaluation of Glaucoma Detection Using Fundus Photographs in Highly Myopic Populations. *Biomedicines*, 12(7), 1394. <https://doi.org/10.3390/biomedicines12071394>
- Demšar, J. (2006). *Statistical Comparisons of Classifiers over Multiple Data Sets*. JMLR.
<https://www.jmlr.org/papers/v7/demsar06a.html>
- Fumero Batista, F. J., Diaz-Aleman, T., Sigut, J., Alayon, S., Arnay, R., & Angel-Pereira, D. (2020). RIM-ONE DL: A Unified Retinal Image Database for Assessing Glaucoma Using Deep Learning. *Image Analysis & Stereology*, 39(3), 161–167. <https://doi.org/10.5566/ias.2346>
- Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1), 3–42. <https://doi.org/10.1007/s10994-006-6226-1>
- Govindharaj, I., Santhakumar, D., Pugazharasi, K., Ravichandran, S., Vijaya Prabhu, R., & Raja, J. (2025). Enhancing glaucoma diagnosis: Generative adversarial networks in synthesized

- imagery and classification with pretrained MobileNetV2. *MethodsX*, 14, 103116. <https://doi.org/10.1016/j.mex.2024.103116>
- Ha, A., & Park, K. H. (2019). Optical Coherence Tomography for the Diagnosis and Monitoring of Glaucoma. *Asia-Pacific Journal of Ophthalmology*. <https://doi.org/10.22608/APO.201902>
- Hassan, T., Akram, M. U., Masood, M. F., & Yasin, U. (2018). *BIOMISA Retinal Image Database for Macular and Ocular Syndromes* (pp. 695–705). https://doi.org/10.1007/978-3-319-93000-8_79
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, 12(1), 5979. <https://doi.org/10.1038/s41598-022-09954-8>
- Hwang, E. E., Chen, D., Han, Y., Jia, L., & Shan, J. (2025). Utilization of Image-Based Deep Learning in Multimodal Glaucoma Detection Neural Network from a Primary Patient Cohort. *Ophthalmology Science*, 5(3), 100703. <https://doi.org/10.1016/j.xops.2025.100703>
- Ilesanmi, A. E., Ilesanmi, T., & Gbotoso, G. A. (2023). A systematic review of retinal fundus image segmentation and classification methods using convolutional neural networks. *Healthcare Analytics*, 4, 100261. <https://doi.org/10.1016/j.health.2023.100261>
- Ling, X. C., Chen, H. S.-L., Yeh, P.-H., Cheng, Y.-C., Huang, C.-Y., Shen, S.-C., & Lee, Y.-S. (2025). Deep Learning in Glaucoma Detection and Progression Prediction: A Systematic Review and Meta-Analysis. *Biomedicines*, 13(2), 420. <https://doi.org/10.3390/biomedicines13020420>
- Meethal, N. S. K., Sisodia, V. P. S., George, R., & Khanna, R. C. (2024). Barriers and Potential Solutions to Glaucoma Screening in the Developing World: A Review. *Journal of Glaucoma*, 33(8S), S33–S38. <https://doi.org/10.1097/IJG.0000000000002404>
- Ministerio de Salud. (2021). *Minsa: Más del 50 % de los pacientes que tiene glaucoma no sabe que lo padece*. Ministerio de Salud. https://www.gob.pe/institucion/minsa/noticias/346283-minsa-mas-del-50-de-los-pacientes-que-tiene-glaucoma-no-sabe-que-lo-padece?utm_source
- Owusu-Adjei, M., Ben Hayfron-Acquah, J., Frimpong, T., & Abdul-Salaam, G. (2023). Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems. *PLOS Digital Health*, 2(11), e0000290. <https://doi.org/10.1371/journal.pdig.0000290>
- Realini, T., McMillan, B., Gross, R. L., Devience, E., & Balasubramani, G. K. (2021). Assessing the Reliability of Intraocular Pressure Measurements Using Rebound Tonometry. *Journal of Glaucoma*, 30(8), 629–633. <https://doi.org/10.1097/IJG.0000000000001892>
- Roboflow. (2026). *GlaucomaData Computer Vision Model*. Universe Roboflow. <https://universe.roboflow.com/classification-8wwxf/glaucomadata/dataset/1>
- Saha, S., Vignarajan, J., & Frost, S. (2023). A fast and fully automated system for glaucoma detection using color fundus photographs. *Scientific Reports*, 13(1), 18408. <https://doi.org/10.1038/s41598-023-44473-0>
- Schuster, A. K., Erb, C., Hoffmann, E. M., Dietlein, T., & Pfeiffer, N. (2020). The Diagnosis and Treatment of Glaucoma. *Deutsches Ärzteblatt International*. <https://doi.org/10.3238/arztebl.2020.0225>
- Sharma, P., Takahashi, N., Ninomiya, T., Sato, M., Miya, T., Tsuda, S., & Nakazawa, T. (2025). A

- hybrid multi model artificial intelligence approach for glaucoma screening using fundus images. *Npj Digital Medicine*, 8(1), 130. <https://doi.org/10.1038/s41746-025-01473-w>
- Shi, D., Zhang, W., Yang, J., Huang, S., Chen, X., Xu, P., Jin, K., Lin, S., Wei, J., Yusufu, M., Liu, S., Zhang, Q., Ge, Z., Xu, X., & He, M. (2025). A multimodal visual–language foundation model for computational ophthalmology. *Npj Digital Medicine*, 8(1), 381. <https://doi.org/10.1038/s41746-025-01772-2>
- Shoukat, A., Akbar, S., Hassan, S. A., Iqbal, S., Mehmood, A., & Ilyas, Q. M. (2023). Automatic Diagnosis of Glaucoma from Retinal Images Using Deep Learning Approach. *Diagnostics*, 13(10), 1738. <https://doi.org/10.3390/diagnostics13101738>
- Sidhu, Z., & Mansoori, T. (2024). Artificial intelligence in glaucoma detection using color fundus photographs. *Indian Journal of Ophthalmology*, 72(3), 408–411. https://doi.org/10.4103/IJO.IJO_613_23
- Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *Computer Vision and Pattern Recognition*, 1. <https://doi.org/https://doi.org/10.48550/arXiv.1409.1556>
- Świerczyński, H., Pukacki, J., Szczęsny, S., Mazurek, C., & Wasilewicz, R. (2025). Application of machine learning techniques in GlaucomAI system for glaucoma diagnosis and collaborative research support. *Scientific Reports*, 15(1), 7940. <https://doi.org/10.1038/s41598-025-89893-2>
- Tan, M., & Le, Q. (2019). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. Proceedings of the 36th International Conference on Machine Learning, PMLR. <https://proceedings.mlr.press/v97/tan19a.html>
- Tham, Y.-C., Li, X., Wong, T. Y., Quigley, H. A., Aung, T., & Cheng, C.-Y. (2014). Global Prevalence of Glaucoma and Projections of Glaucoma Burden through 2040. *Ophthalmology*, 121(11), 2081–2090. <https://doi.org/10.1016/j.ophtha.2014.05.013>
- Zhou, Z.-H., & Feng, J. (2019). Deep forest. *National Science Review*, 6(1), 74–86. <https://doi.org/10.1093/nsr/nwy108>