Universidad Autónoma de Querétaro
Facultad de Ingeniería
Maestría en Ciencias (Inteligencia Artíficial)

CLASIFICACIÓN AUTOMÁTICA DE DEGENERACIÓN MACULAR ASOCIADA A LA EDAD EN IMÁGENES TOMOGRAFÍAS DE COHERENCIA ÓPTICA MEDIANTE APRENDIZAJE PROFUNDO.

TESIS

Que como parte de los requisitos para obtener el Grado de

Maestro en Ciencias (Inteligencia Artíficial)

Presenta:

Oliverio Castillo Rocha

Dirigido por:

Mtro. Oliver Jonathan Quintana Quintana

Mtro. Oliver Jonathan Quintana Quintana
Presidente
Dr. Saúl Tovar Arriaga
Secretario
Dr. Gendry Alfonso Francia
Vocal
Dr. Alberto Hernández Almada
Suplente
Dr. Andras Takacs
Suplente

Centro Universitario, Querétaro, Qro.
Mayo, 2026
México

Dedicatoria

Dedico este trabajo a mi familia, por su apoyo incondicional, su paciencia y la confianza que depositaron en mí durante cada etapa de este proceso. Su esfuerzo, motivación y acompañamiento fueron fundamentales para alcanzar esta meta.

A mis profesores y asesores, por compartir sus conocimientos, orientación y experiencia académica, contribuyendo significativamente a mi formación profesional y personal.

Finalmente, dedico esta tesis a todas las personas que, de una u otra manera, me impulsaron a seguir adelante y a no rendirme ante las dificultades.

Agradecimientos

Expreso mi más sincero agradecimiento al SECIHTI por los recursos otorgados para el desarrollo del presente proyecto de investigación. El apoyo brindado fue fundamental para la realización de este trabajo, permitiendo contar con las condiciones académicas y técnicas necesarias para su correcta ejecución.

A la Universidad Autónoma de Querétaro, agradezco profundamente el respaldo institucional, la formación académica recibida y el entorno académico que hizo posible culminar satisfactoriamente mis estudios de maestría. La calidad docente, el acompañamiento académico y el compromiso con la investigación científica fueron pilares esenciales durante este proceso.

De manera especial, extiendo mi reconocimiento y gratitud a mi comité sinodal por su orientación, retroalimentación crítica y acompañamiento constante a lo largo de este trabajo. Sus observaciones, sugerencias y apoyo académico contribuyeron significativamente a fortalecer la calidad científica del estudio y a consolidar mi formación como profesional e investigador.

Resumen

La degeneración macular asociada a la edad (DMAE) es una de las principales causas de pérdida visual irreversible en adultos mayores, y su diagnóstico oportuno mediante tomografía de coherencia óptica (OCT) puede verse limitado por el tiempo requerido para el análisis manual y la variabilidad entre especialistas. En este contexto, el presente trabajo tiene como objetivo evaluar el desempeño de modelos de aprendizaje profundo para la clasificación automática de imágenes OCT en tres categorías: NORMAL, DRUSEN y CNV. Se propone una arquitectura híbrida basada en CSWin-Transformer como backbone, sobre la cual se integran dos estrategias multiescala: Feature Pyramid Network (FPN) y Atrous Spatial Pyramid Pooling (ASPP), con el fin de determinar cuál ofrece una mejor representación de los patrones retinianos asociados a la DMAE. La metodología incluye un pipeline de preprocesamiento específico para imágenes OCT, compuesto por padding reflectivo, recorte guiado por región de interés, redimensionamiento con conservación de proporción y normalización z-score en la región válida, así como la aplicación de aumentos de datos únicamente durante el entrenamiento. Los experimentos se realizaron utilizando el conjunto público Labeled Retinal OCT Dataset, separando previamente un conjunto de prueba externo a nivel de paciente y aplicando validación cruzada estratificada y agrupada de 5 folds sobre el conjunto restante, con el objetivo de evitar fuga de información y obtener una evaluación robusta. El desempeño se midió mediante exactitud, precisión, recall, Macro-F1, Weighted-F1 y curvas ROC One-vs-Rest con sus respectivos valores de AUC. Los resultados muestran que la configuración CSWin-Transformer + ASPP en su entrenamiento base alcanza el mejor desempeño global, con una exactitud de 0.9440 $\pm$ 0.0073, un Macro-F1 de 0.9433 $\pm$ 0.0069 y un Macro-AUC de 0.9799 $\pm$ 0.0014. Asimismo, se observa que el fine-tuning aporta mejoras limitadas en esta variante, mientras que en FPN sí genera incrementos más notorios sin superar a ASPP. En conclusión, la integración de un Transformer jerárquico con ASPP representa una alternativa eficaz y robusta para la clasificación automática de imágenes OCT en el contexto de la DMAE.

Palabras clave: DMAE, OCT, aprendizaje profundo, CSWin-Transformer.

Abstract

Age-related macular degeneration (AMD) is one of the leading causes of irreversible vision loss in older adults, and its timely diagnosis through optical coherence tomography (OCT) may be limited by the time required for manual analysis and inter-specialist variability. In this context, the objective of this work is to evaluate the performance of deep learning models for the automatic classification of OCT images into three categories: NORMAL, DRUSEN, and CNV. A hybrid architecture based on a CSWin-Transformer backbone is proposed, integrating two multiscale strategies: Feature Pyramid Network (FPN) and Atrous Spatial Pyramid Pooling (ASPP), in order to determine which provides a better representation of retinal patterns associated with AMD. The methodology includes a preprocessing pipeline specifically designed for OCT images, consisting of reflective padding, region-of-interest guided cropping, aspect ratio-preserving resizing, and z-score normalization within the valid region, as well as the application of data augmentation exclusively during training. Experiments were conducted using the public Labeled Retinal OCT Dataset, first separating an external test set at the patient level and then applying stratified and grouped 5-fold cross-validation on the remaining data to prevent information leakage and ensure a robust evaluation. Performance was assessed using accuracy, precision, recall, Macro-F1, Weighted-F1, and One-vs-Rest ROC curves with their corresponding AUC values. The results show that the CSWin-Transformer + ASPP configuration in its base training achieves the best overall performance, with an accuracy of 0.9440 ± 0.0073, a Macro-F1 score of 0.9433 ± 0.0069, and a Macro-AUC of 0.9799 ± 0.0014. Furthermore, fine-tuning provides limited improvements for this variant, while it yields more noticeable gains for FPN without surpassing ASPP. In conclusion, the integration of a hierarchical Transformer with ASPP represents an effective and robust approach for the automatic classification of OCT images in the context of AMD.

Keywords: AMD, OCT, deep learning, CSWin-Transformer.

Índice general

Índice de figuras

Índice de tablas

Abreviaturas

AMD	Age-related Macular Degeneration
ASPP	Atrous Spatial Pyramid Pooling
AUC	Area Under the Curve
BN	Batch Normalization
CAM	Class Activation Mapping
CNN	Convolutional Neural Network
CNV	Choroidal Neovascularization
CPU	Central Processing Unit
CSWin	Cross-Shaped Window Transformer
CV	Cross-Validation
DL	Deep Learning
DMAE	Degeneración Macular Asociada a la Edad
EPR	Epitelio Pigmentario de la Retina
FN	False Negatives
FP	False Positives
FPN	Feature Pyramid Network
FT	Fine-Tuning
GABA	Gamma-Aminobutyric Acid
GAP	Global Average Pooling
GPU	Graphics Processing Unit
Grad-CAM	Gradient-weighted Class Activation Mapping
IA	Inteligencia Artificial
IoU	Intersection over Union
LR	Learning Rate
ML	Machine Learning
OCT	Optical Coherence Tomography
OvR	One-vs-Rest
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristic
ROI	Region of Interest
RPE	Retinal Pigment Epithelium
TN	True Negatives
TP	True Positives
VEGF	Vascular Endothelial Growth Factor

CAPÍTULO 1

Introducción.

El desarrolllo tecnológico ha modificado de gran manera la manera en que se analizan, interpretan y utilizan los datos clínicos en medicina. El análisis de estos datos y las técnicas de imagen de alta resolución han adquirido un papel muy importante para optimizar la toma de decisiones y mejorar la eficiencia en los servicios de salud. Dentro de este escenario, la inteligencia artificial se ha consolidado como una herramienta de apoyo para tareas como la clasificación, la interpretación de estudios médicos y la detección temprana de distintas patologías.

La Degeneración Macular Asociada a la Edad (DMAE) es una de las principales causas de pérdida de visión en personas mayores de 60 años y representa un gran reto para el diagnóstico oportuno. Se estima que actualmente existen alrededor de 196 millones de personas afectadas en el mundo, y que esta cifra podría superar los 288 millones para el año 2040 (Wong et al., 2014). La Organización Mundial de la Salud (World Health Organization, 2019) la reconoce como una de las principales causas de discapacidad visual y ceguera legal en adultos mayores, mientras que el *National Eye Institute* (National Eye Institute, 2022) la clasifica como la enfermedad más común que afecta la mácula en este grupo de edad.

Desde el punto de vista clínico, la DMAE suele distinguirse en dos formas principales. La forma seca o atrófica se relaciona con la presencia de drusas y con cambios degenerativos de evolución progresiva, mientras que la forma húmeda o neovascular se asocia con la aparición de neovascularización coroidea y exudación. En el conjunto de datos empleado en este trabajo, estas manifestaciones se reflejan de manera específica en las etiquetas DRUSEN, vinculada a hallazgos característicos de la forma seca, y CNV, asociada a la forma húmeda; adicionalmente, se incluye la categoría NORMAL, correspondiente a estudios sin hallazgos patológicos relevantes. Por ello, este trabajo se enfoca en la clasificación automática de patrones estructurales observables en imágenes OCT relacionados con dichas manifestaciones, más que en el diagnóstico clínico integral de la enfermedad. En este contexto, la Tomografía de Coherencia Óptica (OCT) constituye una herramienta fundamental, ya que permite visualizar drusas, alteraciones del epitelio pigmentario de la retina y neovascularización coroidea (Drexler & Fujimoto, 2008; MedlinePlus, 2022).

La inteligencia artificial es un área de las ciencias computacionales orientada al desarrollo de sistemas capaces de ejecutar tareas que requieren inferencia, aprendizaje o toma de decisiones. Se

concentra en la creación de máquinas con la capacidad de aprender, procesar datos, tomar decisiones y resolver problemas de forma autónoma, emulando las capacidades cognitivas humanas (Russell & Norvig, 2020).

Durante la última década, la Inteligencia Artificial ha tenido un gran desarrollo utilizando técnicas que ya eran conocidas, como por ejemplo las redes neuronales convolucionales, residuales, etc., las cuales se han podido aplicar con mayor facilidad gracias al uso de las nuevas GPUs que permiten un mejor procesamiento. Tal y como se discute en Russell y Norvig, 2020, las aplicaciones de la IA van desde comprender y procesar el lenguaje hablado hasta vehículos autónomos, asistencia en diagnósticos médicos y algoritmos de predicción utilizados en plataformas digitales, entre otras.

El Deep Learning es parte del Machine Learning, y parte de una gran cantidad de datos; mediante capas de procesamiento se busca que el algoritmo aprenda y pueda realizar tareas para las que fue entrenado, como la identificación de imágenes, el reconocimiento de patrones, el procesamiento de lenguaje o la predicción de valores (Goodfellow et al., 2016).

Aunque la interpretación clínica de imágenes OCT es fundamental, su análisis manual puede ser demandante y presentar variabilidad entre especialistas, sobre todo cuando las alteraciones son sutiles. Frente a esta situación, la inteligencia artificial permite desarrollar modelos capaces de apoyar la identificación de patrones asociados con la DMAE y de hacer más consistente el proceso de clasificación. La integración del conocimiento clínico y las herramientas computacionales abre la puerta a nuevas posibilidades para un diagnóstico asistido más objetivo, reproducible y escalable.

La presente tesis realiza cuatro aportaciones principales. Primero, propone una arquitectura híbrida para clasificación multiclase de imágenes OCT basada en un backbone jerárquico CSWin-Transformer. Segundo, compara de manera controlada dos estrategias de integración multiescala, FPN y ASPP, para determinar cuál representa mejor los patrones estructurales asociados con DRUSEN y CNV. Tercero, implementa un pipeline de preprocesamiento adaptado a imágenes OCT monocanales, orientado a conservar la región anatómica relevante. Cuarto, evalúa ambas variantes mediante un protocolo experimental más estricto, con separación por paciente, conjunto de prueba externo fijo y validación cruzada estratificada y agrupada de 5 folds, con el fin de obtener resultados más robustos y reproducibles.

1.1. Descripción del Problema.

Uno de los retos más importantes en el análisis de imágenes médicas consiste en interpretar de manera precisa patrones visuales asociados con enfermedad. Este reto incluye tanto el procesamiento de las imágenes como la identificación de patrones relevantes que permitan mejorar la precisión diagnóstica. En el caso de la DMAE, el reconocimiento de anomalías en la retina depende de la capacidad para distinguir estructuras oculares sutiles que pueden presentar variaciones mínimas entre pacientes.

El diagnóstico manual mediante Tomografía de Coherencia Óptica requiere un entrenamiento especializado y está sujeto a la experiencia del especialista, lo que puede derivar en errores de clasificación o diagnósticos tardíos. Asimismo, el entrenamiento de modelos de inteligencia artifi-

cial orientados a la detección o clasificación de enfermedades patologías oculares exige bases de datos amplias y bien etiquetadas (así como validadas) para lograr resultados robustos y confiables. Tal como señala Tham et al. (2014), la clasificación precisa de datos médicos permite anticipar la evolución de enfermedades y mejorar la efectividad de los tratamientos en función de información histórica.

En este contexto, la inteligencia artificial ha cobrado una gran importancia en la medicina moderna, y en particular en la oftalmología, donde se ha utilizado en tareas como la clasificación de imágenes, la detección temprana de enfermedades retinianas, la segmentación de estructuras oculares y el seguimiento de su progresión. Estas herramientas no pretenden sustituir al especialista, sino apoyar su trabajo mediante un análisis más rápido, objetivo y consistente de grandes volúmenes de información médica. Desde esta perspectiva, el desarrollo de modelos automáticos para el análisis de imágenes OCT resulta relevante no sólo por su utilidad clínica, sino también por la necesidad de contar con métodos robustos, reproducibles y evaluados mediante protocolos estrictos que eviten la fuga de información entre pacientes. Por ello, es pertinente estudiar arquitecturas de aprendizaje profundo capaces de representar patrones retinianos complejos y comparar su desempeño bajo un esquema experimental sólido.

En el estado del arte reciente muestra que la clasificación de imágenes OCT para patologías retinianas se ha abordado principalmente mediante redes neuronales convolucionales profundas, modelos preentrenados y estrategias multiescala orientadas a integrar información local y contextual. De manera posterior, distintos trabajos incorporaron mecanismos de atención y, más recientemente, arquitecturas basadas en Transformers, capaces de modelar relaciones espaciales de largo alcance. No obstante, persisten diferencias importantes en los protocolos de evaluación, particularmente en la separación por paciente y en el uso de validación cruzada, lo que dificulta realizar comparaciones justas sobre la robustez real de los modelos. Esta situación evidencia la necesidad de estudiar arquitecturas más recientes bajo esquemas experimentales metodológicamente estrictos y comparables.

1.2. Justificación.

La DMAE representa un problema de salud pública de alcance mundial, ya que provoca un deterioro irreversible de la visión central, afectando la autonomía, la productividad y la calidad de vida de millones de personas. Se estima que actualmente existen cerca de 196 millones de personas afectadas, y que esta cifra podría superar los 288 millones para el año 2040 (Wong et al., 2014). En etapas avanzadas, la enfermedad limita la capacidad para leer, conducir o reconocer rostros, generando un impacto considerable en la independencia funcional y en el bienestar psicológico de los pacientes (National Eye Institute, 2022).

El diagnóstico oportuno de la DMAE es crucial para ralentizar la progresión de la enfermedad y preservar la visión útil. Sin embargo, depende de la observación experta y del acceso a equipos especializados, condiciones que no siempre se cumplen de manera uniforme, especialmente en países de ingresos medios como México y otras regiones de América Latina, donde la densidad de oftalmólogos por habitante es limitada y los servicios de salud enfrentan alta demanda (World

Health Organization, 2019). Este panorama hace necesario incorporar herramientas complementarias que fortalezcan la capacidad diagnóstica en distintos contextos clínicos.

La aplicación de IA al análisis de imágenes médicas ha permitido automatizar tareas diagnósticas y disminuir la variabilidad asociada a la interpretación humana. En oftalmología, su aplicación se ha extendido rápidamente a la detección de retinopatía diabética, glaucoma y DMAE, mostrando resultados comparables al desempeño de expertos humanos (Litjens et al., 2017). Los modelos basados en aprendizaje automático y aprendizaje profundo permiten procesar grandes volúmenes de imágenes, identificar patrones complejos y generar predicciones consistentes, contribuyendo así a una atención médica más eficiente y estandarizada.

Particularmente en la Tomografía de Coherencia Óptica, la IA ha permitido desarrollar sistemas capaces de detectar automáticamente alteraciones estructurales, como drusas o desprendimientos del epitelio pigmentario retiniano, que son indicativos de las formas seca y húmeda de la DMAE (Drexler & Fujimoto, 2008). Estos avances favorecen la creación de sistemas de diagnóstico asistido que mejoran el trabajo del especialista y aumentan la disponibilidad de evaluaciones más confiables incluso en contextos de recursos donde existen limitados.

El objetivo de estas herramientas no es sustituir el criterio médico, sino aportar apoyo adicional al proceso de decisión clínica, esto misma permite optimizar el tiempo de análisis, priorizar casos potencialmente patológicos y facilitar el seguimiento de los pacientes. En este sentido, la colaboración interdisciplinaria entre la medicina y las ciencias computaciones se vuelve esencial para garantizar que las soluciones tecnológicas respondan a necesidades reales del entorno clínico (Russell & Norvig, 2020).

Además de su interés técnico, la incorporación de estas herramientas en oftalmología también plantea beneficios potenciales en términos de acceso y oportunidad diagnóstica. Permite ampliar el alcan2ce de la atención a zonas con menor acceso a especialistas, impulsar la telemedicina y mejorar los programas de salud visual. Además, fomenta la investigación colaborativa entre disciplinas y contribuye al desarrollo de una medicina más predictiva y personalizada (Goodfellow et al., 2016).

Desde una perspectiva científica y tecnológica, este trabajo se justifica en la necesidad de evaluar el potencial de la inteligencia artificial como herramienta complementaria para el diagnóstico asistido de enfermedades oculares. Analizar su aplicación en el contexto de la clasificación automática de imágenes de Tomografía de Coherencia Óptica asociadas a la DMAE permitirá sentar bases para el desarrollo de soluciones más precisas, escalables y adaptadas al contexto clínico nacional.

Este trabajo busca evaluar de manera concreta los beneficios del uso de la inteligencia artificial en la detección de la Degeneración Macular Asociada a la Edad, fortalecer el vínculo entre la tecnología y la práctica médica, y contribuir al propósito general de una medicina más preventiva, accesible y orientada al bienestar visual de la población.

CAPÍTULO 2

Antecedentes

2.1. Anatomía del Ojo.

El ojo es un órgano complejo dentro del cuerpo humano, se ubica en la cavidad orbital y aunque en tamaño es pequeño, proporciona información de uno de los cinco sentidos principales, la vista. Para diagnosticar, monitorizar y controlar padecimientos oculares es crucial conocer su estructura y funcionamiento, su función principal es la de detectar estímulos visuales y enviar información visual al cerebro por medio del nervio óptico donde es procesada para obtener una imagen, actúa de manera similar a una cámara, dado que todas sus estructuras trabajan de manera conjunta. Las estructuras que comprende se describe a continuación. (Khurana, 2023)

La córnea es la capa externa del ojo y tiene 2 funciones, las cuales son, servir como capa protectora del mismo así como es responsable del 75 % del potencial óptico del ojo, esta misma carece de vasos sanguíneos y se alimenta mediante el humor acuoso. Tiene la mayor densidad nerviosa y aproximadamente tiene un grosor de 540 micras promedio. (Khurana, 2023)

El iris es la membrana coloreada y circular del ojo que separa la cámara anterior de la cámara posterior. Tiene una abertura central que realiza la comunicación de las dos cámaras que se llama pupila, principalmente controla la cantidad de luz que ingresa en el ojo. Consta de dos músculos los cuáles son esfinter y dilatador, el primero reduce el tamaño de la pupila (miosis) y el segundo lo incrementa (midriasis) (Khurana, 2023).

El cristalino es una estructura del ojo humano con forma de lente biconvexa está situado tras el iris y delante del humor vítreo. Es transparente, avascular y flexible. Se nutre principalmente del humor acuoso. Su anchura aproximada es de 3.5 mm y su propósito es el enfocar objetos a distintas distancias. Cuando se pierde de manera progresiva la transparencia del cristalino se le denomina catarata y se provoca una pérdida de visión (Khurana, 2023).

La retina es la capa más interna del ojo y constituye el tejido neurosensorial encargado de transformar los estímulos luminosos en señales que recibe el nervio óptico, sin embargo su anatomía en particular se describe posteriormente.

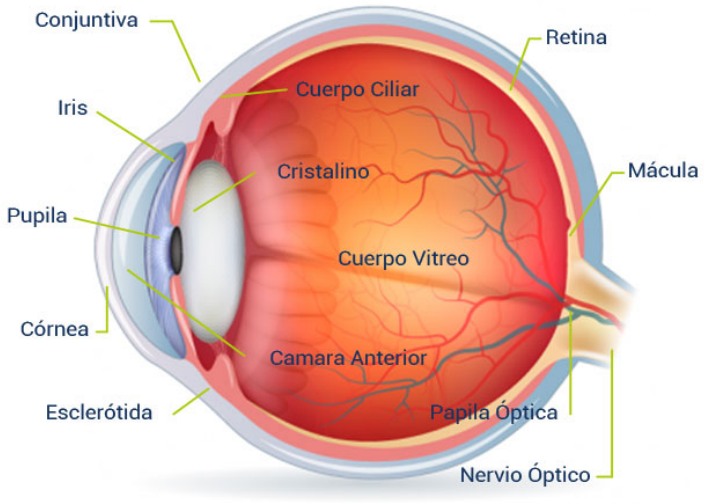


Figura 2.1: Estructura del globo ocular. Tomada de Instituto Diagonal, n.d.

La coroides es una membrana formada por vasos sanguíneos y tejido conectivo. Se encuentra entre la retina y la esclera, su color es oscuro debido a la gran cantidad de melanina que contiene y sirve para evitar que la luz rebote de manera incontrolada dentro del ojo, así como nutrir las capas externas de la retina (Khurana, 2023).

El humor acuoso es un líquido transparente que llena la cámara anterior del ojo, y su función es nutrir y oxigenar las estructuras del globo ocular que no reciben aporte sanguíneo, la córnea y el cristalino (Khurana, 2023).

El gel vítreo es el gel que ocupa la cavidad del globo ocular. Es transparente y está pegado a la retina (Khurana, 2023).

2.1.1. Anatomía de la Retina.

La retina es la capa más profunda del ojo y se extiende desde la salida del nervio óptico hasta el cuerpo ciliar. Su función principal es recibir los rayos de luz y transformarlos en impulsos neuronales que son enviados al cerebro a través del nervio óptico (Alberts et al., 2022; Khurana, 2023). Consta de dos partes fundamentales: la retina neurosensorial interna y el epitelio pigmentario de la retina (RPE). El espacio comprendido entre ambas se conoce como espacio subretiniano; en condiciones normales, las capas permanecen unidas, dejando entre ellas solo un espacio potencial virtual.

Existen dos puntos de referencia topográficos importantes en la retina. La mácula se ubica en

el centro de la retina posterior y es el área donde la visión alcanza su máxima nitidez, dado que contiene una gran densidad de células fotorreceptoras. En su centro se encuentra una depresión poco profunda denominada fóvea central, donde la agudeza visual es máxima. El otro punto de referencia es el disco óptico, localizado aproximadamente a 3 mm de la mácula; corresponde al sitio donde el nervio óptico comienza su trayecto hacia el cerebro. Esta zona carece de células fotorreceptoras, motivo por el cual se le conoce como el "punto ciego" del ojo.

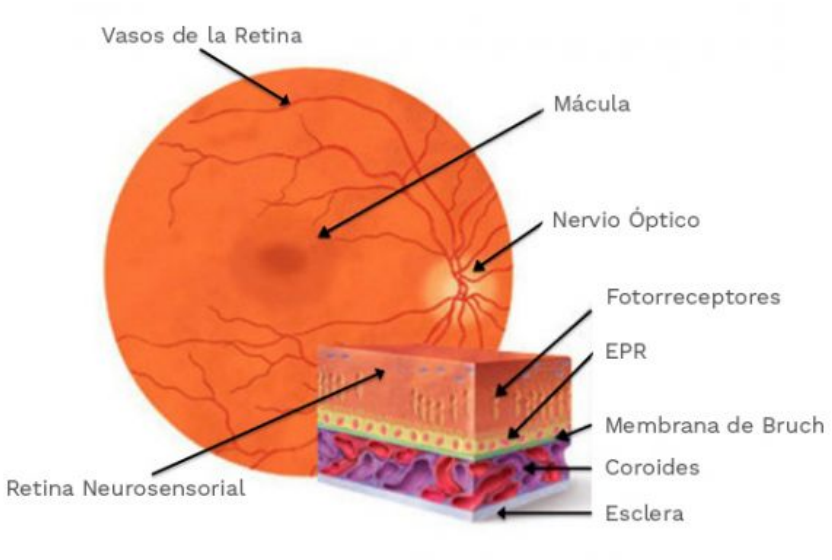


Figura 2.2: Estructura de la retina. Tomada de Ruiz Casas, n.d.

La histología divide la retina en diez capas, dispuestas desde la parte externa hacia la interna (Alberts et al., 2022; Khurana, 2023):

- Capa de fotorreceptores: contiene los segmentos externos de los conos y bastones.
- Membrana limitante externa: separa los segmentos internos y externos de los fotorreceptores y proporciona estabilidad estructural.
- Capa nuclear externa: formada por los núcleos de los fotorreceptores.
- Capa plexiforme externa: región donde los fotorreceptores establecen sinapsis con las células bipolares y horizontales.
- Capa nuclear interna: contiene los núcleos de las células bipolares, horizontales y amacrinas, responsables del procesamiento de la información visual.
- Capa plexiforme interna: zona de conexión entre las células bipolares y las células ganglionares, donde las amacrinas modulan las señales visuales.
- Capa de células ganglionares: formada por los cuerpos celulares de las células ganglionares, cuyos axones se agrupan para constituir el nervio óptico.

- Capa de fibras nerviosas: compuesta por los axones de las células ganglionares que convergen hacia el disco óptico.
- Membrana limitante interna: capa que delimita la retina en su cara interna y la separa del humor vítreo.
- Epitelio pigmentario de la retina (RPE): capa de células pigmentadas que absorbe la luz y nutre la retina neurosensorial.

La retina neurosensorial está compuesta por las primeras nueve capas mencionadas y contiene seis tipos de células especializadas:

- Fotorreceptores: los bastones son células cilíndricas sensibles a la luz de baja intensidad y responsables de la visión escotópica o en escala de grises; los conos, de forma cónica, responden a la luz intensa y son responsables de la visión a color.
- Células bipolares: poseen un axón en un extremo y un árbol dendrítico en el opuesto; establecen sinapsis con los conos y bastones y transmiten la señal hacia las capas más internas de la retina.
- Células ganglionares: son neuronas visuales de segundo orden que reciben información de las células bipolares y amacrinas; sus axones convergen hacia el disco óptico y forman el nervio óptico.
- Células horizontales: realizan sinapsis con los fotorreceptores y las células bipolares, liberando el neurotransmisor GABA para modular la respuesta visual y mejorar el contraste.
- Células amacrinas: se localizan próximas a las células ganglionares y establecen sinapsis con ellas y con las bipolares; regulan la transmisión de las señales visuales, asegurando una respuesta adecuada de las células ganglionares.
- Células de sostén o de Müller: se distribuyen a lo largo de toda la retina neurosensorial; proporcionan soporte estructural y metabólico, y sus prolongaciones forman la membrana limitante interna.

El epitelio pigmentario de la retina es la capa más externa y está formado por células cúbicas con alto contenido de melanina. Se extiende desde el disco óptico hasta la ora serrata. Su función principal es absorber la luz que atraviesa la retina para evitar reflejos internos, además de proporcionar nutrientes a los fotorreceptores y constituir la barrera hematorretiniana, que impide la difusión de moléculas grandes o potencialmente tóxicas hacia la retina neurosensorial.

2.2. Degeneración Macular Asociada a la Edad (DMAE).

La Degeneración Macular Asociada a la Edad (DMAE) es una enfermedad degenerativa de la retina que afecta principalmente a la mácula, una región central responsable de la visión fina y detallada. Es una de las principales causas de pérdida visual irreversible en personas mayores de 60

años, especialmente en países desarrollados, y representa un problema de salud pública creciente a nivel mundial. Según estimaciones globales, la DMAE afecta actualmente a cerca de 196 millones de personas, y se prevé que para el año 2040 esta cifra supere los 288 millones (Wong et al., 2014). La Organización Mundial de la Salud (World Health Organization, 2019) la reconoce como una de las principales causas de discapacidad visual, mientras que el *National Eye Institute* (National Eye Institute, 2022) la considera la enfermedad macular más común asociada al envejecimiento.

La mácula es la zona de la retina donde la agudeza visual alcanza su máximo nivel, debido a la alta densidad de conos que permite distinguir los detalles más finos y los colores. En la DMAE, esta región se ve comprometida por alteraciones estructurales y funcionales del epitelio pigmentario de la retina (EPR), los fotorreceptores y la coroides, lo que conlleva a una pérdida progresiva de la visión central, mientras la periférica permanece generalmente conservada (Lim, 2013).

Existen dos formas clínicas de la enfermedad. La forma seca o atrófica representa aproximadamente el 85 % de los casos, y se caracteriza por la acumulación de drusas —depósitos amarillentos de material extracelular entre el EPR y la membrana de Bruch—, el adelgazamiento progresivo del epitelio pigmentario y la pérdida gradual de fotorreceptores. Por otro lado, la forma húmeda o neovascular, menos frecuente pero más agresiva, se caracteriza por el crecimiento anómalo de vasos sanguíneos procedentes de la coroides (neovascularización coroidea o CNV) que invaden la retina, provocando filtraciones, hemorragias y cicatrices que deterioran rápidamente la visión (MedlinePlus, 2022).

Entre los factores de riesgo más relevantes se incluyen la edad avanzada, antecedentes familiares de DMAE, el tabaquismo, la hipertensión arterial, la exposición prolongada a la luz ultravioleta, una dieta pobre en antioxidantes y el sexo femenino, que presenta una prevalencia ligeramente superior (Khurana, 2023; Lim, 2013). La presencia de drusas blandas o de grandes dimensiones es uno de los principales indicadores del riesgo de progresión hacia formas más avanzadas de la enfermedad.

Desde el punto de vista clínico, los pacientes con DMAE suelen manifestar disminución de la agudeza visual, dificultad para leer, distorsión de las líneas rectas (*metamorfopsias*) y aparición de manchas oscuras en el campo visual central (*escotomas*). El diagnóstico se realiza mediante la exploración del fondo de ojo y el uso de técnicas de imagen como la tomografía de coherencia óptica (OCT), la angiografía fluoresceínica o la autofluorescencia, que permiten identificar cambios estructurales en la retina y el epitelio pigmentario (Drexler & Fujimoto, 2008).

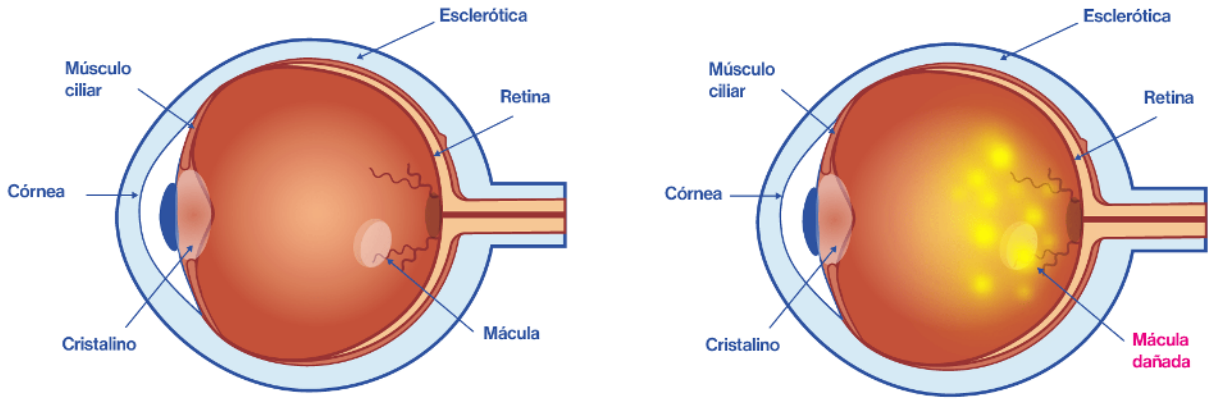


Figura 2.3: Comparativa entre ojo sano y afectado por DMAE, en el segundo caso se observa el daño en la mácula central. Tomada de Oftalvist, n.d.

2.2.1. Fisiopatología de la DMAE.

La fisiopatología de la Degeneración Macular Asociada a la Edad es multifactorial y compleja, implicando procesos de envejecimiento celular, estrés oxidativo, disfunción del epitelio pigmentario de la retina (EPR), inflamación crónica y alteraciones en la perfusión coroidea. Estas interacciones generan un deterioro progresivo de la homeostasis retiniana y comprometen el soporte metabólico de los fotorreceptores.

Con el envejecimiento, el metabolismo del EPR se vuelve menos eficiente, lo que favorece la acumulación de lipofuscina —un pigmento derivado de los restos de los fotorreceptores— en sus células. Al mismo tiempo, se produce un engrosamiento y pérdida de permeabilidad de la membrana de Bruch, lo que dificulta el intercambio de nutrientes y desechos entre la coroides y la retina (Khurana, 2023; Lim, 2013). Este proceso promueve la formación de drusas, acúmulos extracelulares de lípidos, proteínas y material inflamatorio que se depositan entre la membrana de Bruch y el EPR.

En la forma seca, el daño progresivo del EPR y la atrofia del epitelio pigmentario generan una pérdida gradual de fotorreceptores, provocando una reducción lenta pero irreversible de la agudeza visual central. En la forma húmeda, el deterioro de la membrana de Bruch y la liberación de factores angiogénicos, especialmente el *factor de crecimiento endotelial vascular* (VEGF), estimulan la proliferación de vasos coroideos anómalos que invaden el espacio subretiniano. Estos vasos son frágiles y propensos a la filtración de líquido o sangre, lo que genera edema y cicatrices fibrosas que destruyen la arquitectura retiniana (Khurana, 2023; MedlinePlus, 2022).

El estrés oxidativo desempeña un papel central en la patogenia de la DMAE. La exposición continua a la luz y al oxígeno favorece la formación de radicales libres en los fotorreceptores y el EPR. Cuando la capacidad antioxidante natural disminuye, estos radicales provocan daño lipídico, proteico y del ADN celular, contribuyendo a la muerte celular y la disfunción metabólica. Además, se ha descrito la activación del complemento y la presencia de inflamación crónica de bajo grado

como mecanismos adicionales que agravan el proceso degenerativo (Khurana, 2023; Lim, 2013).

En conjunto, estos procesos llevan a una alteración irreversible de las capas retinianas centrales, con pérdida de los fotorreceptores y disfunción del EPR. Esta secuencia patológica explica la progresión clínica de la DMAE desde la etapa inicial con presencia de drusas, pasando por la atrofia geográfica en la forma seca, hasta la neovascularización activa en la forma húmeda. Los avances en técnicas de imagen como la OCT han permitido visualizar de manera precisa estos cambios estructurales, consolidando su uso como herramienta diagnóstica y de seguimiento (Drexler & Fujimoto, 2008; National Eye Institute, 2022).

2.3. Técnicas de Imagen para la Evaluación de la DMAE.

En el diagnóstico y seguimiento de la Degeneración Macular Asociada a la Edad (DMAE) se emplean diversas técnicas de imagen que permiten evaluar las alteraciones estructurales en la retina, la mácula y el epitelio pigmentario. Estas herramientas no solo contribuyen a la detección temprana de la enfermedad, sino que también facilitan el monitoreo de su progresión y la respuesta al tratamiento (Khurana, 2023; Lim, 2013). A continuación, se describen las principales técnicas utilizadas en oftalmología para la evaluación de la DMAE.

Fotografía de fondo de ojo: es una técnica convencional que utiliza una cámara oftálmica para obtener imágenes en color del fondo del ojo, incluyendo la retina, la mácula y el nervio óptico. Permite identificar signos característicos como la presencia de drusas, alteraciones pigmentarias y atrofia del epitelio pigmentario. Sin embargo, no proporciona información en profundidad sobre las capas retinianas, por lo que su uso suele complementarse con otras técnicas más avanzadas.

Angiografía fluoresceínica: consiste en la inyección intravenosa de fluoresceína sódica y la posterior captura secuencial de imágenes para evaluar la circulación retinocoroidea. Es fundamental para detectar la neovascularización coroidea (CNV) característica de la DMAE húmeda y para valorar fugas o áreas de hipoperfusión. Aunque ha sido una técnica de referencia, su carácter invasivo y la posibilidad de reacciones adversas limitan su uso sistemático (Lim, 2013; National Eye Institute, 2022).

Angiografía con verde de indocianina: similar a la fluoresceínica, pero utiliza un colorante con mayor afinidad por los vasos coroideos. Permite visualizar con mayor detalle la vasculatura subretiniana y detectar membranas neovasculares ocultas. Se usa principalmente para complementar el diagnóstico en casos complejos o en combinación con la Tomografía de Coherencia Óptica.

Autofluorescencia del fondo de ojo: esta técnica aprovecha la fluorescencia natural de la lipofuscina acumulada en el epitelio pigmentario de la retina. Es útil para detectar daño celular temprano y monitorizar la progresión de la atrofia geográfica en la DMAE seca.

Tomografía de coherencia óptica: se basa en la interferometría de baja coherencia para generar cortes tomográficos de alta resolución de la retina. Permite observar en detalle las capas retinianas, identificar drusas, desprendimientos del epitelio pigmentario, neovascularización y atrofias localiza-

das. Su carácter no invasivo, su rapidez y su elevada resolución axial la han convertido en la técnica de referencia para el diagnóstico y seguimiento de la DMAE (Drexler & Fujimoto, 2008; National Eye Institute, 2022).

OCT-Angiografía (OCT-A): es una extensión reciente de la OCT que permite visualizar el flujo sanguíneo sin necesidad de agentes de contraste. Facilita la identificación de membranas neovasculares y la cuantificación de la perfusión retinocoroidea, siendo especialmente útil en la DMAE húmeda.

En este trabajo, el uso de la OCT es fundamental, ya que constituye la herramienta principal para analizar los cambios morfológicos en la retina y la mácula asociados a la enfermedad. A continuación, se describe con mayor detalle su principio de funcionamiento, aplicaciones y relevancia en el diagnóstico asistido mediante inteligencia artificial.

2.3.1. Tomografía de Coherencia Óptica (OCT).

La Tomografía de Coherencia Óptica, ilustrada en la Fig. 2.4, es una técnica no invasiva que permite obtener imágenes de alta resolución de las secciones transversales de los tejidos oculares. Ofrece un amplio rango de aplicaciones en el ámbito de la salud visual, ya que proporciona información estructural detallada de la retina y la mácula (Drexler & Fujimoto, 2008).

- Evaluación de condiciones oculares: la OCT y sus algoritmos de segmentación se emplean para analizar y diagnosticar diversas enfermedades de la retina, entre ellas la degeneración macular asociada a la edad, la retinopatía diabética y el edema macular. Permite medir el grosor y la morfología de las capas retinianas, lo que facilita la detección temprana de alteraciones estructurales.

- Seguimiento de enfermedades: en patologías crónicas como la DMAE húmeda, la OCT permite realizar un seguimiento preciso de la evolución de las lesiones y valorar la respuesta a los tratamientos, especialmente en terapias antiangiogénicas.

- Investigación clínica y farmacológica: en estudios experimentales y ensayos clínicos, la segmentación de OCT se utiliza para evaluar la eficacia de tratamientos farmacológicos y terapias emergentes dirigidas a enfermedades retinianas.

- Cirugía asistida por OCT: en cirugía oftalmológica, la OCT puede emplearse como herramienta intraoperatoria para guiar procedimientos de retina o segmento anterior, mejorando la precisión quirúrgica y la evaluación del resultado postoperatorio.

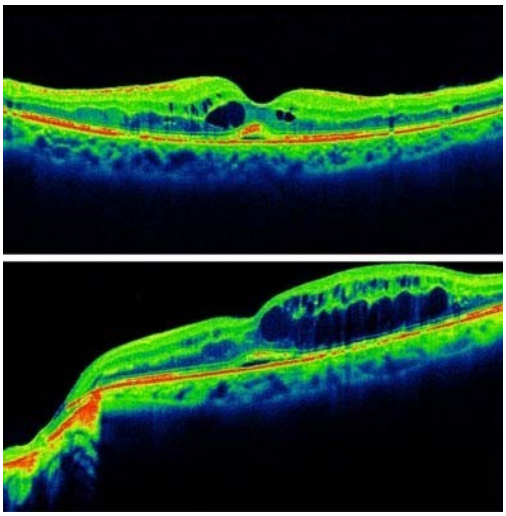


Figura 2.4: Tomografía de Coherencia Óptica. Tomada de Scanner Vizcaya, 2017.

El principio básico de funcionamiento de la OCT, mostrado en la Fig. 2.5, se basa en la interferometría de baja coherencia. En este método, una fuente de luz se divide en dos haces: uno se dirige hacia el tejido ocular y el otro hacia un espejo de referencia. La luz reflejada por las diferentes capas de la retina se combina con la del haz de referencia, generando un patrón de interferencia que varía según la profundidad del tejido analizado.

El procesamiento de estos patrones permite calcular las distancias relativas entre las capas retinianas y reconstruir imágenes tridimensionales del área de estudio. Los resultados se presentan en forma de secciones transversales o tomogramas, donde es posible observar el grosor y la morfología de cada capa de la retina con una resolución del orden de micras (Drexler & Fujimoto, 2008).

2.3.1.1. OCT en el Dominio del Tiempo vs. OCT en el Dominio Espectral.

La principal diferencia entre la tomografía de coherencia óptica en el dominio del tiempo (TD-OCT) y en el dominio espectral (SD-OCT) radica en la forma en que se registran las señales de interferencia de baja coherencia y en la tecnología empleada para procesar las imágenes. Estos factores determinan la velocidad de adquisición, la resolución y la precisión de los datos obtenidos.

La TD-OCT fue la primera generación de esta tecnología y marcó un avance decisivo en la obtención de imágenes transversales de la retina. Sin embargo, con los progresos en óptica y procesamiento digital, la SD-OCT se ha consolidado como el estándar actual en oftalmología, al ofrecer mayor velocidad, resolución y estabilidad en el diagnóstico temprano de enfermedades retinianas, incluida la degeneración macular asociada a la edad (Drexler & Fujimoto, 2008; Khurana, 2023).

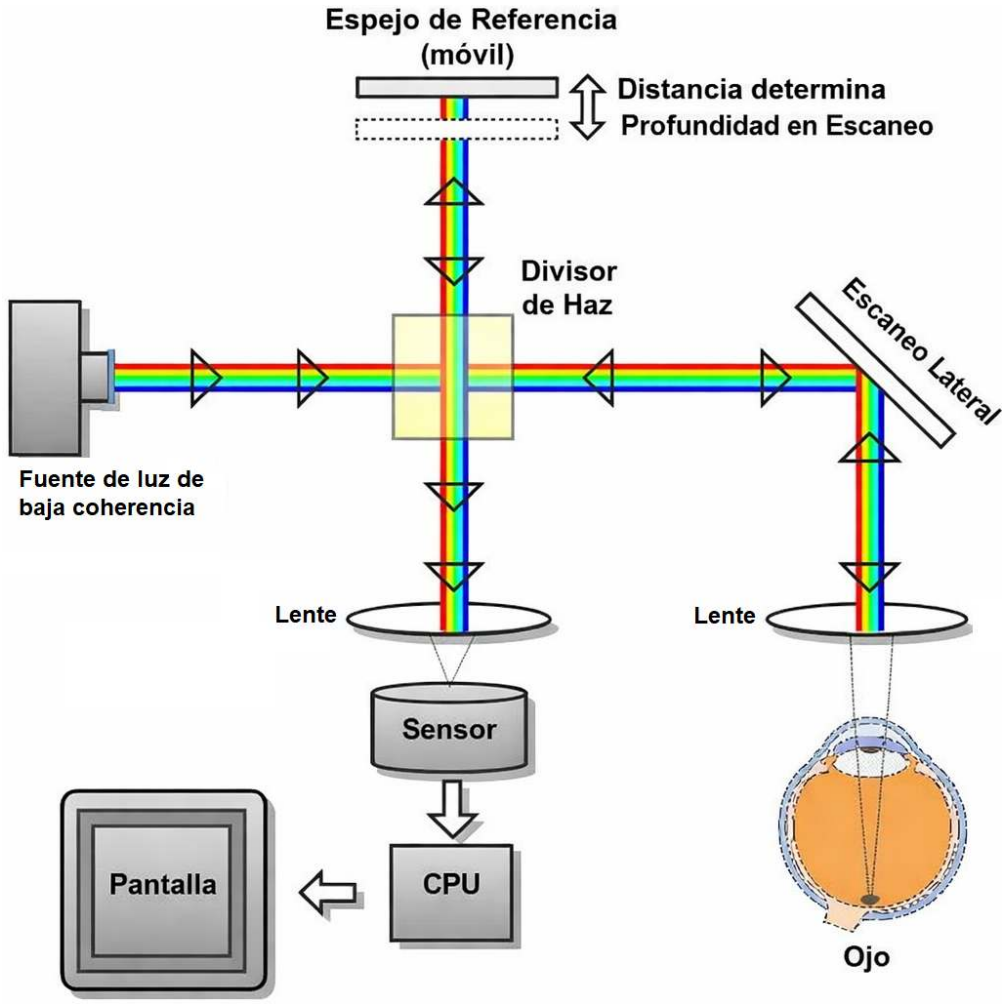


Figura 2.5: Esquema de OCT. Tomada de Cabaleiro et al., 2019.

Tabla 2.1: Comparación entre TD-OCT y SD-OCT

Característica	**TD-OCT**	**SD-OCT**
Método de captura	Basado en el tiempo (espejo móvil)	Basado en frecuencia (espectrómetro o cámara CCD)
Velocidad de captura	∼100 líneas de barrido por segundo	∼1000 líneas de barrido por segundo
Resolución axial	10–15 µm	5–7 µm
Precisión y profundidad	Moderada	Alta
Aplicación	Diagnóstico básico	Diagnóstico detallado y cuantitativo

2.4. Indicadores de DMAE en OCT.

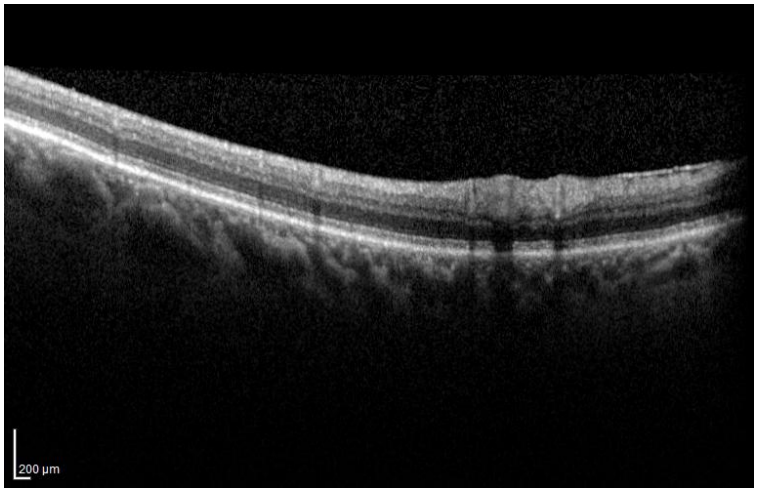
La tomografía de coherencia óptica se ha consolidado como una herramienta esencial en el diagnóstico, clasificación y monitoreo de la Degeneración Macular Asociada a la Edad (DMAE). Su capacidad para generar cortes tomográficos de alta resolución permite visualizar con precisión las alteraciones estructurales de la mácula y el epitelio pigmentario de la retina (EPR), lo que facilita la detección precoz de lesiones y la evaluación de la respuesta terapéutica (Arsalan et al., 2022; Drexler & Fujimoto, 2008; Lim, 2013; National Eye Institute, 2022; Shi et al., 2023).

Los hallazgos más característicos observables en imágenes OCT en pacientes con DMAE incluyen los siguientes:

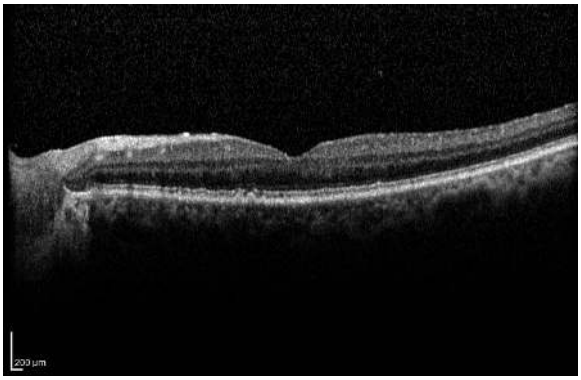
- Drusas: representan el signo más temprano y frecuente de la DMAE seca. Se observan como elevaciones focales e irregulares del EPR sobre la membrana de Bruch. Su tamaño y reflectividad varían según su composición y grado de evolución; las drusas blandas, de bordes difusos, se asocian a un mayor riesgo de progresión hacia estadios avanzados (Khurana, 2023; Lim, 2013).
- Alteraciones del epitelio pigmentario de la retina (EPR): comprenden zonas de hiperreflectividad e hiporreflectividad que evidencian daño celular, acumulación de lipofuscina o desprendimiento del EPR (PED). En la OCT, estos desprendimientos se visualizan como elevaciones bien definidas, homogéneas o irregulares cuando existe neovascularización subyacente (Arsalan et al., 2022).
- Atrofia geográfica: corresponde a la pérdida del EPR y de los fotorreceptores, acompañada de adelgazamiento retiniano y aumento de la reflectividad coroidea. La OCT revela discontinuidades de la banda del EPR y reducción significativa del espesor foveal, características de la fase avanzada de la DMAE seca (Ledesma-Carbayo et al., 2023; Shi et al., 2023).
- Neovascularización coroidea (CNV) es la manifestación distintiva de la forma húmeda. En la OCT se aprecia como una elevación irregular del EPR con material subretiniano o intrarretiniano de alta reflectividad. La CNV suele asociarse a fluido exudativo, hemorragias o fibrosis subretiniana. Su identificación es fundamental para establecer el inicio o la modificación de la terapia anti-VEGF (Kadir et al., 2023; Lim, 2013; National Eye Institute, 2022).
- Fluido intrarretiniano y subretiniano: se manifiesta como espacios hipo o anecoicos entre las capas retinianas o debajo del EPR. Estos hallazgos reflejan actividad exudativa y permiten cuantificar el grado de respuesta terapéutica mediante el análisis longitudinal de tomogramas consecutivos (Arsalan et al., 2022; Kadir et al., 2023).
- Desprendimiento del EPR y de la retina neurosensorial: ocurre cuando el EPR o las capas internas se separan de la membrana de Bruch o entre sí por acumulación de fluido o tejido neovascular. En la OCT se observa como una elevación en forma de cúpula o serosa con bordes bien delimitados (Drexler & Fujimoto, 2008; Lim, 2013).
- Cambios morfológicos en la mácula: incluyen alteraciones en la arquitectura de las capas internas, irregularidades en la fóvea y variaciones del espesor macular central. Estos parámetros,

cuantificados mediante mapas de grosor, se correlacionan con la función visual y la severidad de la enfermedad (Ledesma-Carbayo et al., 2023).

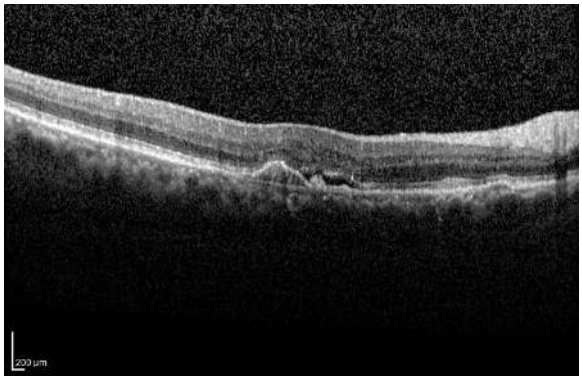
En conjunto, estos indicadores permiten una caracterización detallada del estado anatómico de la retina, facilitando la distinción entre DMAE seca y húmeda, la evaluación de la actividad neovascular y el monitoreo de la eficacia terapéutica, de modo que se ilustra de manera comparativa en la Figura 2.6. La precisión y reproducibilidad de la OCT han consolidado su papel como herramienta diagnóstica de referencia en oftalmología moderna (Arsalan et al., 2022; Drexler & Fujimoto, 2008; National Eye Institute, 2022).



(a) OCT normal, estructura retiniana conservada, con capas bien definidas y sin signos de DMAE.



(b) DMAE seca (Drusas): elevaciones focales del epitelio pigmentario y depósitos subretinianos.



(c) DMAE húmeda (CNV): presencia de neovascularización coroidea y fluido subretiniano.

Figura 2.6: Comparativa OCT de retina normal y de las dos formas de DMAE. En la parte superior se muestra la arquitectura retiniana normal, mientras que en la parte inferior se observan los patrones característicos de la b) DMAE seca (drusas) y de la a) DMAE húmeda (CNV con exudación). Imágenes del conjunto de datos Sotoudeh-Paima et al., 2023.

00

2.5. Aprendizaje Automático (Machine Learning).

El aprendizaje automático es una rama de la inteligencia artificial orientada a construir modelos capaces de identificar regularidades en los datos y utilizarlas para resolver tareas específicas, En términos generales, su propósito es reconocer patrones útiles en los datos y emplearlos para apoyar procesos de clasificación, predicción o toma de decisiones. (Russell & Norvig, 2020).

Así mismo, también se puede lograr que se encuentren relaciones y tendencias, utilizando esta información para realizar tareas específicas. La idea fundamental es que, a medida que se enfrentan a más datos, las máquinas se vuelven más capaces y precisas en las tareas asignadas.

En el contexto del Machine Learning, es importante distinguir entre dos enfoques:

2.5.0.1. Aprendizaje supervisado.

En el aprendizaje supervisado, el modelo se entrena con ejemplos cuya salida esperada ya es conocida, de modo que aprende una relación entre las entradas y sus etiquetas. Es decir, cada entrada de datos se asocia con una etiqueta. El algoritmo aprende a relacionar las entradas con las etiquetas y, una vez entrenado, puede predecir etiquetas para nuevas entradas basándose en lo que ha aprendido. Este método se utiliza en tareas como la clasificación y la regresión, donde se intenta predecir una categoría o un valor numérico (Bishop, 2006).

2.5.0.2. Aprendizaje no supervisado.

A diferencia del caso anterior, el aprendizaje no supervisado trabaja con datos sin etiqueta y busca descubrir estructuras internas presentes en ellos. En este caso, el algoritmo busca patrones y estructuras en los datos, encontrando correlaciones sin conocer cuáles son los resultados deseados. Esto es útil para tareas como la clusterización, donde el objetivo es agrupar datos similares, o para la reducción de la dimensionalidad, que permite simplificar datos complejos (Bishop, 2006).

2.6. Aprendizaje Profundo (Deep Learning).

El aprendizaje profundo se apoya en redes neuronales con múltiples capas que permiten construir representaciones progresivamente más abstractas a partir de los datos de entrada. A diferencia de los enfoques tradicionales, donde las características deben diseñarse manualmente, los modelos de Deep Learning aprenden automáticamente múltiples niveles de abstracción mediante la composición sucesiva de transformaciones no lineales. Este paradigma ha demostrado un rendimiento sobresaliente en tareas de visión por computadora, procesamiento de lenguaje natural y análisis de datos médicos (Goodfellow et al., 2016).

Durante el ntrenamiento para el aprendizaje profundo, se alimenta a la red neuronal con una cantidad de datos de entrada, tales como imágenes, texto o sonido, y se ajustan los pesos y sesgos de las neuronas en cada capa mediante la propagación hacia atrás, con el fin de aprender patrones

y características a diferentes niveles de la red. A medida que los datos pasan a través de las capas, las representaciones adquieren nuevas características y, por lo tanto, interpretaciones distintas (Krizhevsky et al., 2012).

El aprendizaje profundo ha revolucionado el campo de la inteligencia artificial y se ha aplicado en una amplia variedad de tareas, como el reconocimiento de patrones en imágenes, el procesamiento de lenguaje natural, la traducción de lenguaje, la generación de texto y el procesamiento de voz. Es extremadamente poderoso y versátil, pero, como se discute en Goodfellow et al., 2016, también puede requerir grandes cantidades de datos de entrenamiento y recursos computacionales para obtener un rendimiento óptimo. Su capacidad para capturar y representar información compleja lo convierte en una herramienta fundamental en la actualidad para abordar problemas de inteligencia artificial y aprendizaje automático de alta complejidad.

Mediante la utilización de redes neuronales artificiales con múltiples capas de procesamiento, se logra un funcionamiento conjunto para extraer, transformar y normalizar características de los datos antes de llevarlos a las capas subsecuentes en el proceso (Goodfellow et al., 2016).

El proceso de capas tiene tres fases principales:

- Capas de entrada: responsables de aprender características simples a partir de los datos de entrada. Pueden detectar bordes, patrones o colores básicos.

- Capas ocultas: ubicadas entre las capas de entrada y salida. Aprenden características más complejas y abstractas, construyendo representaciones de niveles superiores basadas en las características extraídas en las capas de entrada.

- Capas de salida: generan las predicciones o resultados finales del modelo. Utilizan las características aprendidas en las capas ocultas para realizar tareas específicas, como clasificación, regresión o generación de texto.

El Deep Learning es especialmente adecuado para tareas que involucran grandes cantidades de datos, ya que estas redes neuronales tienen la capacidad de aprender representaciones de datos altamente complejas. Algunas de las aplicaciones más destacables son:

- Predicciones: mediante el aprendizaje con los datos proporcionados, al recibir entradas nuevas puede predecir una respuesta.

- Procesamiento del Lenguaje Natural: se utiliza para realizar traducción automática, generación de texto o comprensión de voz.

- Visión por Computadora: el Deep Learning se emplea en aplicaciones de visión por computadora, como la detección de objetos, el reconocimiento facial, la identificación de patrones y la segmentación de imágenes, incluyendo las imágenes OCT.

2.6.0.1. ¿Qué son las redes neuronales?.

Las redes neuronales son modelos computacionales diseñados para aprender relaciones entre entradas y salidas a partir de ejemplos. Su diseño se inspira de forma conceptual en la organización de las neuronas biológicas, aunque su implementación matemática responde a operaciones algebraicas y funciones de activación (Fig. 2.7).

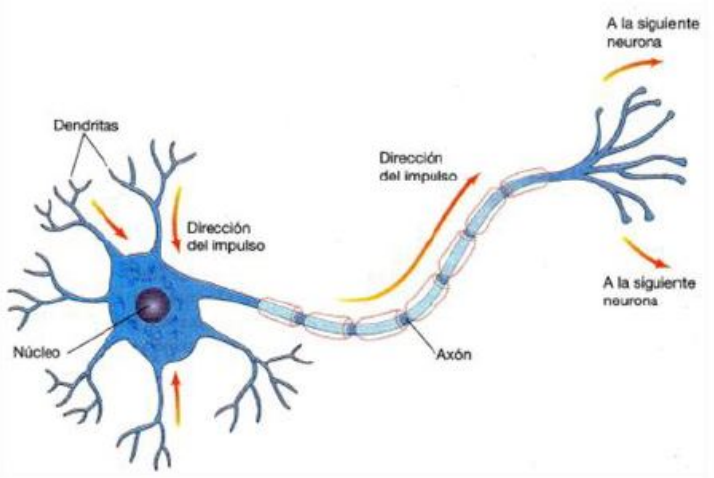


Figura 2.7: Neurona Humana. Alberts et al., 2022.

La neurona se estimula por estímulos externos y, cuando alcanza un cierto umbral, se activa, transmitiendo una señal hacia el axón (Alberts et al., 2022). Este principio biológico sirve de base para la organización y el funcionamiento de las redes neuronales artificiales. Estas se utilizan en el ámbito de la informática y el aprendizaje automático para realizar tareas específicas, como identificar patrones, categorizar datos, procesar lenguaje natural o tomar decisiones.

Tal y como se discute en Krizhevsky et al., 2012, las redes se conforman por un conjunto de unidades interconectadas que se asemejan a las células cerebrales. Estas unidades se agrupan en diferentes niveles: una etapa de inicio, niveles intermedios (cuya densidad y cantidad dependerán del problema a resolver) y un nivel de salida. Cada enlace entre estas unidades tiene un valor asignado que regula la interacción (peso de conexión).

El funcionamiento de una red neuronal se basa en el procesamiento de información a través de estas conexiones y niveles. La información fluye desde la fase de inicio, pasa por las capas intermedias y finalmente llega al nivel de salida, donde se produce la respuesta o predicción deseada, tal y como se muestra en la Figura 2.8.

Las redes neuronales pueden aprender de los datos mediante un proceso de entrenamiento. Durante este proceso, los valores de conexión se ajustan teniendo en cuenta ejemplos previos de entrada y salida conocidos, lo que permite que la red adquiera la capacidad de identificar patrones y ejecutar tareas específicas. Este ajuste se realiza mediante algoritmos de optimización, como el método de descenso del gradiente.

Un aspecto destacado de las redes neuronales es su capacidad para aprender de manera no

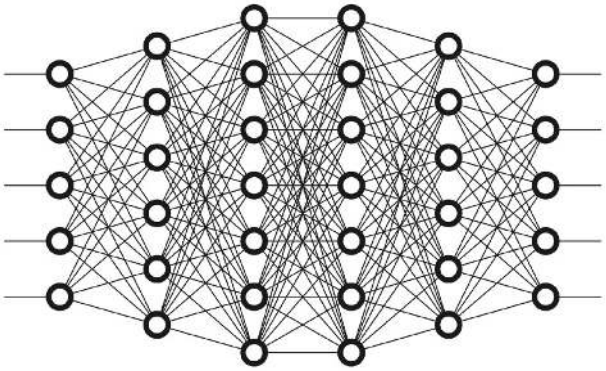


Figura 2.8: Estructura genérica de una Red Neuronal. Goodfellow et al., 2016.

lineal y representar relaciones complejas en los datos. Esto las convierte en herramientas versátiles, aplicables en una amplia gama de contextos, desde la interpretación de imágenes y el procesamiento de lenguaje hasta la predicción de datos y la toma de decisiones.

2.6.1. Tipos de Redes Neuronales.

El aprendizaje profundo abarca una amplia gama de arquitecturas diseñadas para procesar diferentes tipos de datos y relaciones. En general, las redes neuronales se pueden clasificar en cuatro grandes grupos: *feedforward*, convolucionales, recurrentes y basadas en atención. Cada una presenta particularidades estructurales y operativas que las hacen adecuadas para distintas tareas en distintas áreas del conocimiento.

2.6.1.1. Redes Feedforward (FNN).

Las redes *feedforward* constituyen la forma más elemental de red neuronal artificial, caracterizándose por un flujo unidireccional de la información desde la capa de entrada hasta la capa de salida, sin la presencia de ciclos o mecanismos de retroalimentación interna. En este tipo de arquitectura, cada capa recibe las activaciones de la capa anterior y las transforma mediante una combinación de operaciones lineales y no lineales, generando representaciones progresivamente más abstractas.

Desde el punto de vista matemático, una red *feedforward* puede describirse como una composición sucesiva de funciones paramétricas:

$$y = f_L(f_{L-1}(\dots f_2(f_1(x))\dots)) \tag{2.1}$$

donde x representa la entrada, y la salida, y cada f_i corresponde a una transformación aprendible compuesta típicamente por una operación afín seguida de una función de activación no lineal. Esta estructura garantiza que el grafo computacional sea acíclico, permitiendo que la propagación

hacia adelante (*forward propagation*) y el cálculo del gradiente mediante retropropagación (*back-propagation*) se realicen de manera eficiente.

Su estructura puede estar conformada por capas densas o completamente conectadas, en las cuales cada neurona de una capa se conecta con todas las neuronas de la siguiente. Entre los modelos más representativos de este paradigma se encuentran el Perceptrón Multicapa (MLP) y los Autoencoders. Asimismo, el enfoque *feedforward* incluye arquitecturas más profundas y estructuralmente complejas que preservan el flujo directo de información, como las redes convolucionales profundas.

Dentro de estas variantes modernas se encuentran las Redes Neuronales Residuales (ResNet), que introducen conexiones de salto (*skip connections*) con el objetivo de facilitar el flujo del gradiente durante el entrenamiento y mitigar problemas asociados al incremento de la profundidad (He et al., 2016a). A pesar de estas modificaciones estructurales, dichas arquitecturas mantienen la naturaleza *feedforward*, ya que el grafo computacional continúa siendo acíclico y la información fluye en dirección entrada–salida.

En general, las redes *feedforward* son ampliamente utilizadas para tareas de clasificación y regresión, y constituyen la base conceptual de la mayoría de las arquitecturas modernas de *Deep Learning* (Goodfellow et al., 2016). Su capacidad para aproximar funciones altamente no lineales, respaldada por el teorema de aproximación universal, las convierte en una herramienta fundamental para el modelado de problemas complejos en distintos dominios.

2.6.1.2. Redes Neuronales Convolucionales (CNN).

Las Redes Neuronales Convolucionales (CNN) son un tipo de red diseñada específicamente para el análisis de datos que exhiben una estructura cuadricular, como imágenes o secuencias temporales. Este tipo de redes son particularmente efectivas en tareas relacionadas con el procesamiento de imágenes y han demostrado un rendimiento sobresaliente en distintas aplicaciones. Su estructura general se visualiza en la Figura 2.9 (Krizhevsky et al., 2012).

Como se discute en Krizhevsky et al., 2012, las redes neuronales convolucionales tienen distintas capas, descritas a continuación:

- Capas convolucionales: constituyen el núcleo de una CNN y son el lugar donde ocurre la mayor parte del cálculo. Requieren distintos componentes, tales como los datos de entrada, un filtro (*kernel*) y un mapa de características. El detector de características básicamente es una matriz de pesos que representa una sección de la imagen; este se aplica a un área determinada y se obtiene la operación producto punto entre los píxeles de entrada y el filtro. El resultado se almacena en una matriz de salida, repitiéndose el proceso hasta cubrir toda la imagen. Como resultado, se obtiene una serie de productos punto de la operación convolución, conocidos como mapas de características. Esto permite que la red detecte patrones locales y características específicas. Existen tres hiperparámetros que afectan el volumen de salida y deben definirse antes de iniciar el entrenamiento:
    - Número de filtros: determina la profundidad de la salida. Si se aplican tres filtros distintos, se obtendrán tres mapas de características diferentes, con una profundidad de tres.

- Stride: es la distancia o número de píxeles que el núcleo se desplaza sobre la matriz de entrada. Aunque los valores de *stride* iguales o superiores a dos son poco frecuentes, un valor mayor produce una salida de menor tamaño.

- Capas de agrupación (*pooling*): se utilizan para reducir la resolución espacial de las características extraídas en las capas convolucionales. Esto contribuye a la reducción de la cantidad de parámetros, al mismo tiempo que conserva las características más significativas.

- Capas completamente conectadas: al final de la arquitectura de la CNN, típicamente se incorporan capas totalmente conectadas que procesan las características extraídas y generan una salida, como una clasificación o una regresión.

Este tipo de redes muestra alta eficacia en la clasificación de objetos en imágenes, detección de patrones y procesamiento de imágenes médicas. Esta última es el fin de este trabajo. Su capacidad para aprender representaciones jerárquicas de características visuales ha tenido un gran impacto en campos como la conducción autónoma, el diagnóstico médico y la detección de objetos en tiempo real para aplicaciones de seguridad.

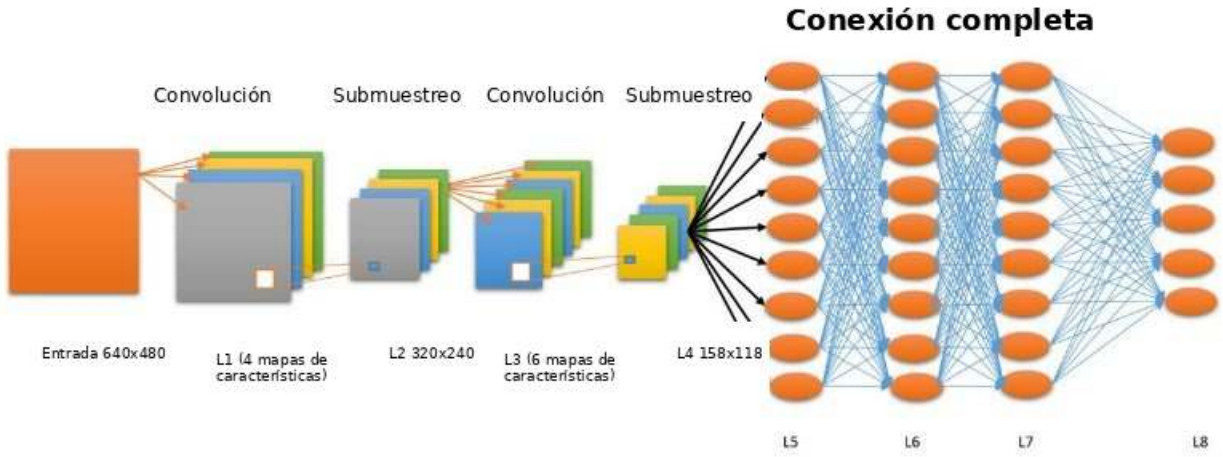


Figura 2.9: Red neuronal convolucional. De la Rosa, 2024.

Dentro de la familia de redes convolucionales, se han propuesto múltiples extensiones para mejorar su capacidad de capturar información a diferentes escalas espaciales. Entre ellas destacan la *Feature Pyramid Network* (FPN) y el módulo *Atrous Spatial Pyramid Pooling* (ASPP).

La FPN, propuesta por Lin et al., 2017, permite combinar características de distintas resoluciones para generar mapas multiescala, siendo ampliamente utilizada en tareas de detección y segmentación. Por su parte, el módulo ASPP, introducido por Chen, Papandreou et al., 2018 en DeepLabv3, emplea convoluciones dilatadas con diferentes tasas de dilatación para capturar contextos espaciales de múltiples tamaños sin aumentar significativamente la complejidad computacional. Estas estrategias resultan particularmente útiles en imágenes médicas como las de OCT, donde las estructuras anatómicas pueden variar considerablemente en tamaño y forma, dependiendo incluso del equipo con el que fueron adquiridas.

2.6.1.3. Redes Neuronales Recurrentes (RNN).

Las redes neuronales recurrentes (RNN) están diseñadas para procesar datos secuenciales, como series temporales, señales o texto. A diferencia de las redes feedforward, las RNN incorporan bucles internos que permiten que la información de pasos anteriores influya en el procesamiento de los actuales, dotándolas de una forma de "memoria", de modo que se ilustra en la Figura 2.10.

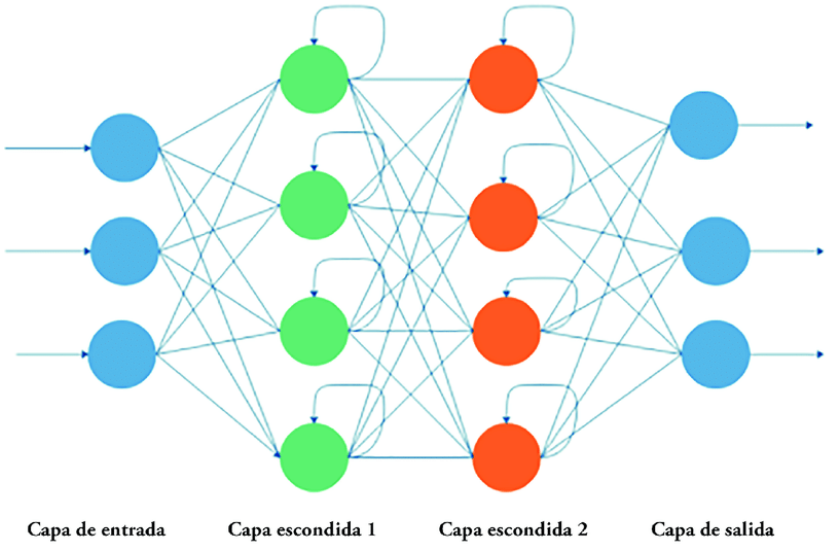


Figura 2.10: Red neuronal recurrente. Franco y Ramos, 2019.

Entre sus variantes más utilizadas se encuentran las Long Short-Term Memory (LSTM) y las Gated Recurrent Units (GRU), que resuelven los problemas de gradientes desaparecientes comunes en las RNN básicas. Estas redes son ampliamente usadas en reconocimiento de voz, predicción temporal y análisis de secuencias biológicas, aunque en el campo de la visión por computadora han sido progresivamente reemplazadas por los modelos basados en atención.

2.6.1.4. Redes Basadas en Atención (Transformers).

Los Transformers son un tipo muy especial de red neuronal que cambió las tareas de procesamiento de lenguaje natural (*Natural Language Processing*, NLP) y otras áreas de la IA. Introducida por Vaswani et al., 2017 en el artículo "*Attention Is All You Need*", marcó un punto de inflexión gracias a la forma en que manejan la información secuencial, eliminando completamente los mecanismos recurrentes utilizados previamente.

Desde un punto de vista estructural, los Transformers son redes *feedforward* profundas, ya que la información fluye en una sola dirección a través de capas apiladas. Sin embargo, su incorporación del mecanismo de autoatención multi-cabeza les otorga un comportamiento y una capacidad de modelado contextual que los distingue de las redes tradicionales, motivo por el cual se consideran una familia aparte dentro del aprendizaje profundo.

A diferencia de los modelos secuenciales tradicionales, como las RNN o las LSTM, los Transformers se basan en el mecanismo de atención para procesar secuencias completas de manera paralela.

Esto les permite no depender del orden estrictamente iterativo de los datos, facilita el paralelismo durante el entrenamiento y capta mejor las relaciones de largo alcance, evitando los problemas de desvanecimiento o explosión del gradiente asociados a arquitecturas recurrentes profundas.

El mecanismo de atención puede entenderse como la capacidad de "enfocarse" en las partes más relevantes de una secuencia cuando se procesan o generan datos. Así, en una oración con varias palabras, el modelo aprende cuáles aportan más información para la tarea específica (traducción, resumen, pregunta-respuesta, etc.). Desde un punto de vista matemático, este proceso se formaliza mediante la proyección lineal de la representación de entrada en tres espacios distintos.

Sea una secuencia representada como una matriz:

$$X \in \mathbb{R}^{n \times d}$$

donde n es el número de tokens y d la dimensión del embedding. A partir de esta representación se obtienen tres matrices:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V$$

donde W_Q, W_K, W_V son matrices de parámetros aprendibles.

Consta de tres elementos principales:

- Key (Clave): representa la información con la que se comparará la consulta.
- Query (Consulta): indica qué está buscando el modelo en la secuencia.
- Value (Valor): es la información que se transmite cuando la Key y la Query coinciden de forma relevante.

De este modo, se calculan pesos que indican cuánto aporta cada Key a la Query mediante un producto punto escalado:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

El factor $\frac{1}{\sqrt{d_k}}$ actúa como mecanismo de estabilización numérica, evitando que los valores crezcan excesivamente cuando la dimensión del espacio proyectado es grande. El resultado es una combinación ponderada de los valores V, donde los pesos reflejan la relevancia contextual entre tokens.

En la práctica, el Transformer no utiliza una única operación de atención, sino múltiples cabezas de atención en paralelo (*Multi-Head Attention*). Cada cabeza aprende relaciones distintas en subespacios diferentes, y sus resultados se concatenan para formar una representación más rica y expresiva.

La arquitectura general del Transformer se divide en dos bloques principales: el *Encoder* y el *Decoder*.

El Encoder toma una secuencia de entrada, convierte cada palabra (o subpalabra) en un vector de *embedding*, incorpora información posicional mediante codificaciones senoidales o aprendibles, y aplica varias capas compuestas por: (1) autoatención multi-cabeza, (2) conexiones residuales, (3) normalización por capas (*Layer Normalization*) y (4) redes completamente conectadas (*feed-forward*). El resultado final es una representación interna contextualizada que resume lo más relevante de la secuencia de entrada (Vaswani et al., 2017).

El Decoder utiliza la representación producida por el Encoder para generar la secuencia de salida. Trabaja de manera autoregresiva, generando un token (palabra o subpalabra) a la vez, teniendo en cuenta lo que ya ha producido. Para ello emplea atención enmascarada (que impide ver posiciones futuras) y atención cruzada sobre la representación proveniente del Encoder.

La ventaja de separar ambos bloques es que el Encoder se especializa en entender la información de entrada, mientras que el Decoder se encarga de generar la salida de forma coherente y ordenada. Este diseño modular, junto con la capacidad de modelar dependencias globales con complejidad cuadrática respecto al número de tokens, ha permitido que los Transformers demuestren un rendimiento sobresaliente en tareas como la traducción automática, la generación de texto y la clasificación de oraciones (Brown et al., 2020); (Devlin et al., 2018).

2.7. Arquitecturas basadas en aprendizaje profundo.

El uso de DL ha transformado el análisis de imágenes médicas y, en particular, en OCT, dada su alta resolución y el detalle que ofrecen sobre las capas de la retina, es fundamental contar con modelos capaces de extraer representaciones robustas para detectar signos de patologías como la DMAE, la retinopatía diabética y el glaucoma. A continuación, se revisan arquitecturas relevantes para este ámbito médico.

2.7.1. Arquitecturas basadas en CNN.

Las CNN se emplean como *backbones*, es decir, extractores jerárquicos de características que transforman la imagen en representaciones de mayor nivel. Sobre estos extractores se acoplan cuellos multiescala y una cabeza de clasificación. A continuación se explican las arquitecturas CNN más utilizadas y pertinentes dado el contexto de la tarea abordada en esta tesis, destacando su funcionamiento y el equilibrio entre capacidad representacional y costo computacional.

VGG (2014). Arquitectura profundamente apilada de bloques *conv* 3×3–*ReLU* (Rectified Linear Unit, $f(x) = \max(0, x)$) seguidos de *max-pooling* que reduce a la mitad la resolución tras cada bloque; conforme disminuye $H \times W$, se duplican los canales (p. ej., 64→128→256→512). La cabeza original es totalmente conectada, aunque en la práctica se sustituye por *Global Average Pooling* (GAP, que es el promedio espacial por canal de $H \times W$ a 1×1) + capa lineal para reducir el *número de parámetros* y el riesgo de sobreajuste. Su practicidad facilita el análisis, pero la pila de conv 3×3 y las FC finales hacen que los parámetros y los *FLOPs* (operaciones en coma flotante por inferencia) sean elevados (Simonyan & Zisserman, 2015).

ResNet (2015). Introduce *bloques residuales* con *atajos* (identidad o proyección 1×1) que suman la entrada a la salida del bloque, estabilizando el gradiente en redes muy profundas. Sus variantes con *bottlenecks* 1×1–3×3–1×1 (reducir–procesar–expandir) contienen el cómputo y los *FLOPs* sin sacrificar capacidad.ResNet suele requerir menos parámetros efectivos para alcanzar altas precisiones, y su diseño modular facilita la tarea de acoplar cuellos (FPN/ASPP) y una cabeza de clasificación (He et al., 2016b).

DenseNet (2017). En este caso capa recibe la concatenación de todas las salidas previas dentro de un bloque denso, promoviendo la reutilización de características. El ancho crece con una tasa

de crecimiento k, y entre bloques se insertan *transition layers* (1×1 + *avg-pooling*) que comprimen y reducen la resolución. Esta conectividad densa ocasiona buena precisión con menos parámetros que arquitecturas a su nivel, aunque aumenta el consumo de memoria intermedia y puede elevar los *FLOPs* si el crecimiento k y la resolución inicial son altos (Huang et al., 2017).

Inception / Xception (2015–2017). Inception sigue el principio *split–transform–merge*: división del flujo en ramas paralelas con filtros 1×1, 3×3 y 5×5 (a menudo factorizados en $1\times N$ y $N\times 1$) y fusiona por concatenación; los 1×1 reducen dimensionalidad antes de operaciones costosas, controlando parámetros y *FLOPs*. Xception toma la misma idea y la incrementa usando *depthwise separable convolutions* (convolución por canal seguida de punto a punto 1×1), que desacoplan mezcla espacial y de canales, disminuyendo drásticamente los *FLOPs* sin perder capacidad multiescala; incorpora además atajos residuales para estabilidad (Chollet, 2017; Szegedy et al., 2015).

MobileNet (V1–V3, 2017–2019). Estas arquitecturas están diseñadas para optimizar para recursos limitados: V1 emplea sistemáticamente *depthwise separable conv* para reducir *FLOPs* y parámetros; V2 introduce *inverted residuals* (expandir–profundizar con *depthwise*–comprimir) con *linear bottlenecks* que preservan información en espacios de baja dimensión; V3 combina búsqueda de arquitectura con atenciones ligeras (por ejemplo autoatención) y activaciones eficientes. Con cabeza GAP + lineal, mantiene la relación precisión–costo, especialmente en resoluciones moderadas (A. Howard et al., 2019; A. G. Howard et al., 2017; Sandler et al., 2018).

EfficientNet (2019). Propone un *escalado compuesto* que ajusta coordinadamente profundidad, ancho y resolución a partir de una base (B0) diseñada por búsqueda neuronal. Sus bloques *MBConv* (inverted residual con *depthwise* y atención *Squeeze-and-Excitation*) ofrecen alta relación precisión/*FLOPs*; las variantes B0–B7 permanecen sobre una frontera de Pareto favorable. Es notable que en este tipo de arquitecturas, usar GAP + lineal como cabeza permite aprovechar su eficiencia en parámetros y *FLOPs* para usarse de backbone moderno en tareas de clasificación (Tan & Le, 2019).

Tabla 2.2: Arquitecturas CNN

Arquitectura	**Funcionamiento**	**Ventaja típica**
VGG	Bloques conv 3×3 + max-pooling	Base estable y simple; muchos parámetros y FLOPs
ResNet	Conexiones residuales (atajos)	Entrenamiento profundo fiable; buen baseline general
DenseNet	Conexiones densas intra-bloque	Reutiliza características, menos parámetros; mayor memoria intermedia
Inception/Xception	Ramas paralelas y factorizar conv	Multiescala en el bloque; mayor complejidad de diseño
MobileNet	Convoluciones separables y cuellos invertidos	Muy eficiente (móvil/embebido); ligera caída de precisión
EfficientNet	Escalado compuesto + MBConv + SE	Mejor relación precisión/eficiencia; backbone popular y ligero

2.7.1.1. Segmentación.

La tarea de segmentación en OCT consiste en delimitar automáticamente estructuras anatómicas y/o para generar máscaras píxel a píxel que posibilitan medir espesores, áreas y cambios en la longitud. A diferencia de la clasificación (etiqueta global de la imagen), la segmentación produce un mapa detallado por píxel.

Por ejemplo U-Net es una arquitectura *encoder–decoder* con *skip connections* que preservan detalle fino al reconstruir la resolución original, y constituye un estándar de facto en segmentación biomédica, incluida la OCT (Ronneberger et al., 2015). Alcance: la segmentación se presenta únicamente como contexto conceptual y técnico; forma parte del modelo experimental implementado. Este trabajo se centra en clasificación de OCT .

2.7.1.2. Módulos multiescala/piramidales (cuellos): FPN y ASPP.

Las características en OCT (drusas, exudados, bordes de membranas) aparecen en distintas escalas dimensionales. Por ello, es común acoplar al *backbone* cuello multiescala que integre información de varias resoluciones antes de la cabeza de clasificación. Dos variantes ampliamente utilizadas son *Feature Pyramid Networks (FPN)* (Lin et al., 2017) y *Atrous Spatial Pyramid Pooling (ASPP)* (Chen, Zhu et al., 2018; Chen et al., 2017). Recientemente la combinación de FPN sobre un backbone CNN con OCT ha mostrado mejoras consistentes frente a sus contrapartes que no la incluyen. (Sotoudeh-Paima et al., 2022).

FPN: fusión piramidal top–down con conexiones laterales. Con un *backbone* jerárquico (VGG/ResNet/EfficientNet) que produce mapas en múltiples niveles $\{C_2, C_3, C_4, C_5\}$, FPN construye una *pirámide de rasgos* $\{P_2, \ldots, P_5\}$ combinando:

1. **Proyecciones laterales** 1×1 de cada C_ℓ para igualar canales.
2. **Camino top–down** con *upsampling* (típicamente ×2) desde $P_{\ell+1}$.
3. **Fusión** por suma: $P_\ell = \text{proj}(C_\ell) + \text{upsample}(P_{\ell+1})$, seguida de un conv 3×3 de "suavizado" que reduce aliasing.

Así, cada P_ℓ combina la semántica profunda (capas altas) con la localización fina (capas bajas). En la clasificación de OCT, puede usarse (i) un *cuello* a una sola escala (P_3 o concatenar $\text{GAP}(P_2 \| P_3 \| P_4 \| P_5)$) o (ii) una *cabeza* que agregue las pirámides con atención ligera antes de la capa final, esto incrementa la robustez a escala, permite reutilización de mapas del backbone y baja el costo adicional (Lin et al., 2017; Sotoudeh-Paima et al., 2022).

ASPP: contexto multiescala con dilataciones paralelas. ASPP agrega contexto sin reducir resolución aplicando convoluciones
textitatrous (dilatadas) en paralelo, con diferentes *rates* $\{r_1, r_2, r_3, \ldots\}$ que separan los filtros y amplían el campo receptivo efectivo. Un bloque típico consta de:

- Un conv 1×1 (base).

- Varias ramas 3×3 con dilataciones $r \in \{6, 12, 18\}$ (ejemplo a 224^2).
- (Opcional) una rama de pooling global + 1×1 para aportar contexto global de la imagen.

Se concatenan las salidas y se proyectan con 1×1. Frente a max-pooling o strides, las dilataciones preservan el detalle espacial, de vital importancia en OCT, donde bordes de capas finas y micro-lesiones se degradan con subsampling agresivo. ASPP funciona como cuello genérico (CNN o Transformer) antes de la cabeza de clasificación, aportando sensibilidad a escalas sin incrementar los FLOPs (Chen, Zhu et al., 2018; Chen et al., 2017)

Tabla 2.3: Comparativa práctica entre FPN y ASPP, y aspectos de implementación en OCT.

	Feature Pyramid Network (FPN)	**Atrous Spatial Pyramid Pooling (ASPP)**
Naturaleza	Fusiona niveles jerárquicos ya existentes del *backbone* mediante conexiones laterales y un camino *top–down*.	Enriquece un único nivel con campos receptivos múltiples usando convoluciones dilatadas paralelas.
Costo computacional	Añade proyecciones 1×1, *upsampling* y conv 3×3 de suavizado.	Añade varias ramas 3×3 con diferentes dilataciones; costo controlado, sin *downsampling* adicional.
Escenario ideal de uso	Cuando el *backbone* produce una jerarquía profunda (ResNet, EfficientNet); útil para integrar semántica y detalle.	Cuando se busca mayor contexto en un solo mapa (C3/P3) o se trabaja con imágenes de menor resolución.
Combinación	Se puede usar para construir $\{P_\ell\}$ jerárquicos.	Frecuentemente se aplica sobre un nivel de la FPN (P_3) antes de la cabeza de clasificación.
	Implementación	
Normalización y activación	Usar BN o GN después de conv 1×1 y 3×3, con activaciones ReLU o SiLU.	
Elección de niveles	En B-scan OCT de 512×512, emplear niveles $\{P_2$–$P_5\}$; en 224×224, centrarse en P_3/P_4 para evitar mapas excesivamente pequeños.	
Dilataciones en ASPP	Ajustar las tasas de dilatación según la resolución: menores para 224^2, mayores para 512^2, evitando el efecto *gridding*.	

2.7.1.3. Atención en CNN: SE, CBAM.

La atención dentro de las CNN es una extensión natural para reforzar la selectividad de las características extraídas por el backbone, adaptando dinámicamente la respuesta de la red según la importancia de cada canal o región espacial analizada. Estos mecanismos anteceden al concepto

de autoatención global de los Transformers, aportando una transición conceptual entre las CNN puras y los modelos híbridos actuales.

Squeeze-and-Excitation. Se propuso por Hu et al. (2018), este módulo introduce una rama de recalibración de canales mediante dos etapas: *squeeze*, que aplica un GAP para resumir la respuesta espacial de cada canal, y después ese pasa vector por dos capas totalmente conectadas con activaciones ReLU y Sigmoide, generando pesos de atención por canal. Al multiplicar estos pesos con el mapa original se refuerzan los canales más informativos, logrando mejoras sistemáticas en tareas de clasificación con un costo computacional bajo.

Convolutional Block Attention Module. El módulo CBAM amplía la idea de SE al combinar atención por canal y por espacio en dos etapas secuenciales (Woo et al., 2018). Tras aplicar una atención de canales similar a SE, incorpora una segunda operación que calcula un mapa espacial de atención a partir de las agregaciones promedio y máxima sobre los canales anteriores, destacando regiones relevantes dentro del plano espacial. Su estructura ligera permite insertarlo fácilmente en arquitecturas como ResNet o DenseNet sin incrementar de gran manera *FLOPs*.

2.7.2. Arquitecturas basadas en Transformers.

Los Transformers, originalmente introducidos para el procesamiento del lenguaje natural por Vaswani et al. (2017), se fundamentan en el mecanismo de *self-attention*, que permite modelar dependencias a largo alcance entre elementos de una secuencia sin recurrir a estructuras recurrentes o convoluciones. Posteriormente este principio se extendió a la visión por computadora, donde las imágenes se reinterpretan como secuencias de tokens que interactúan entre sí mediante atención. De esta manera, los *Vision Transformers* (ViT) reemplazan la operación convolucional local por atención global, capturando relaciones espaciales de manera más explícita y flexible.

La gran diferencia de las CNN es que la recepción crece gradualmente con la profundidad de la red, sin embargo los Transformers tienen la capacidad conexiones directas entre regiones distantes desde las primeras capas. Esto se traduce en una mayor eficiencia para representar patrones estructurales globales, aunque con un costo computacional cuadrático respecto al número de tokens de entrada. Existen diversas variantes que han sido propuestas para reducir dicho costo o introducir jerarquía espacial, buscando un equilibrio entre eficiencia y precisión. Entre ellas se destacan *Swin-Transformer*, *Pyramid Vision Transformer* (PVT) y el *CSWin-Transformer*, este último de gran importancia en este trabajo por su atención eficiente en ventanas cruzadas (Dong et al., 2022).

En el contexto de imágenes OCT, las arquitecturas basadas en Transformers tienen gran potencial, ya que combinan la capacidad de atención global con la posibilidad de preservar detalles locales críticos para la detección de alteraciones retinianas y a su vez identificar patologías. Estas ventajas fundamentan su uso como *backbone* dentro de modelos híbridos, donde se integran con módulos multiescala o piramidales, como FPN o ASPP, para un análisis jerárquico y contextual más completo.

2.7.2.1. ViT y variantes jerárquicas.

El ViT es la primera aplicación directa de los Transformers a la visión por computadora. Este enfoque toma una imagen y la divide en tokens fijos (por ejemplo, 16×16 píxeles), cada uno de los cuales se aplana y se proyecta linealmente a un espacio de *embeddings* (una representación numérica densa y continua de una entidad discreta) que conforma una secuencia de tokens de entrada (Vaswani et al., 2017). Estos tokens se procesan mediante múltiples capas de *multi-head self-attention* (MHSA) y bloques *feed-forward*, análogos a los empleados en modelos de lenguaje. En tareas de clasificación, se incluye un token adicional (*class token*) cuya representación final resume la información global de la imagen. Aunque ViT obtuvo buenos en conjuntos muy grandes como ImageNet-21k, su rendimiento en escenarios con datos limitados fue muy inferior al de las CNN debido a la falta de inductivas espaciales (traslación, localidad).

A fin de mejorar su eficiencia y capacidad de generalización, surgieron variantes jerárquicas que incorporan estructuras piramidales y ventanas locales de atención. El *Pyramid Vision Transformer* (PVT) introduce una reducción progresiva de resolución y un escalado de canales similar a las CNN, permitiendo obtener mapas de características a múltiples niveles jerárquicos. Posteriormente, el *Swin Transformer* propuso un mecanismo de atención en ventanas desplazadas (*shifted windows*), que restringe la atención a regiones locales y la desplaza entre capas para favorecer la comunicación global sin elevar el costo computacional cuadráticamente.

Estas variantes jerárquicas consolidaron la idea de emplear Transformers como *backbone* visual, los cuales son comparables a las redes convolucionales profundas. Modelos más recientes, como el *CSWin-Transformer*, utilizan el diseño de las ventanas de atención mediante una disposición cruzada que captura interacciones horizontales y verticales de manera más eficiente (Dong et al., 2022). La preservación del detalle local y el contexto global es gran punto fuerte del mismo, por lo tanto esta estrategia es adecuada para tareas médicas de alta resolución, como es el caso de la clasificación de imágenes de OCT.

2.7.3. Arquitectura CSWin-Transformer.

Este apartado tendrá un énfasis muy importante y detallado, dado que es el backbone central de este trabajo de investigación como así se detalla.

CSWin-Transformer (Dong et al., 2022) es un *backbone* jerárquico de visión con cuatro etapas de procesamiento. Su distintivo se basa en la autoatención en ventanas con forma de cruz, el cual consta que en cada bloque, la mitad de las cabezas atiende en franjas horizontales y la otra mitad en franjas verticales; luego se concatenan y proyectan. Además, el ancho de franja s_w crece con la profundidad (pequeño al inicio, grande al final). El crecimiento es ocasionado por el paralelismo horizontal/vertical amplía el área de interacción por bloque sin añadir pasos secuenciales extra, mejorando el campo receptivo con coste controlado. El s_w dependiente de la etapa equilibra cobertura y eficiencia: cuando las resoluciones son grandes, s_w pequeño reduce cómputo; cuando son pequeñas, s_w grande aporta más contexto con sobrecosto moderado.

Las OCT muestran estructuras laminares y patrones en ambos ejes del plano de modo que la

atención utilizada por este *backbone* toma esas dependencias dentro del mismo bloque, evitando atención global costosa o cadenas secuenciales. La pirámide de salidas facilita FPN/ASPP antes de la cabeza de clasificación, y LePE aporta robustez cuando varía el tamaño dado el preprocesamiento.

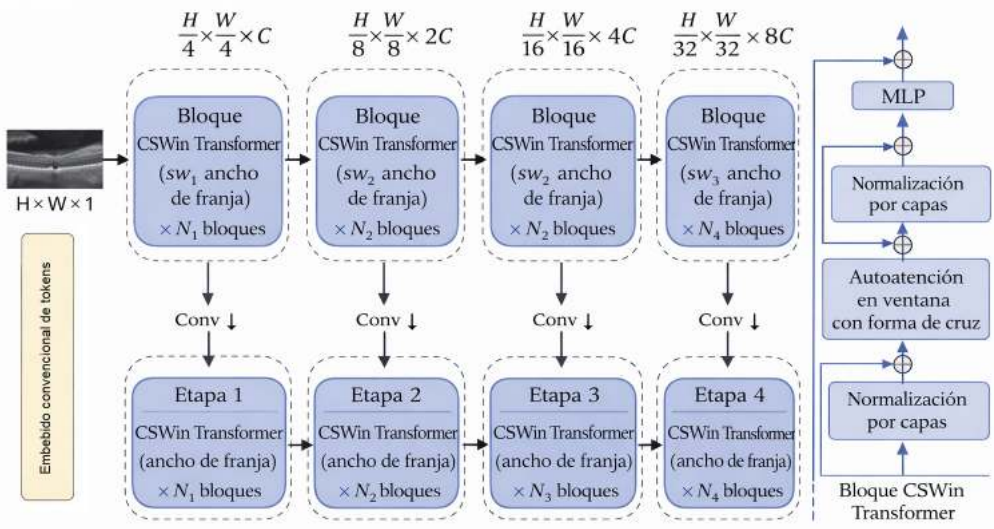


Figura 2.11: Estructura general del CSWin-Transformer con cuatro etapas jerárquicas y bloques de atención en ventanas con forma de cruz. Adaptada de Dong et al. (2022).

2.7.3.1. Mecanismo de atención.

Se utiliza un mecanismo de atención ventana con forma de cruz, de modo que, sea $\boldsymbol{X}_\ell \in \mathbb{R}^{H_\ell \times W_\ell \times C_\ell}$ el tensor de la etapa ℓ. Con un ancho de franja $s_w^{(\ell)}$, se definen dos dominios ortogonales de atención (franjas horizontales y verticales). Las proyecciones por cabeza son:

$$\begin{aligned} \boldsymbol{Q} &= \boldsymbol{X}_\ell \, \boldsymbol{W}_Q, \\ \boldsymbol{K} &= \boldsymbol{X}_\ell \, \boldsymbol{W}_K, \\ \boldsymbol{V} &= \boldsymbol{X}_\ell \, \boldsymbol{W}_V, \end{aligned} \tag{2.2}$$

donde $\boldsymbol{W}_\bullet \in \mathbb{R}^{C_\ell \times d}$ y $d = \dfrac{C_\ell}{h_\ell}$.

La atención restringida a un dominio $\mathcal{D} \in \{\mathrm{H}, \mathrm{V}\}$ se calcula como:

$$\mathrm{Attn}_{\mathcal{D}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top + \mathbf{M}_{\mathcal{D}}}{\sqrt{d}} \right) \mathbf{V}, \tag{2.3}$$

donde $\mathbf{M}_{\mathcal{D}}$ enmascara las posiciones fuera de la franja horizontal o vertical correspondiente. La *ventana con forma de cruz* resulta de concatenar las salidas de ambos dominios (a lo largo de cabezas) y aplicar una proyección lineal final.

En la etapa ℓ, el tensor de características es $\boldsymbol{X}_\ell \in \mathbb{R}^{H_\ell \times W_\ell \times C_\ell}$, donde H_ℓ y W_ℓ son alto y ancho espaciales, y C_ℓ es el número de canales. Para calcular la atención multi–cabeza, se proyecta $\boldsymbol{X}_\ell$ a

tres espacios latentes mediante transformaciones lineales: consultas $\boldsymbol{Q}$, claves $\boldsymbol{K}$ y valores $\boldsymbol{V}$. Estas proyecciones se obtienen multiplicando $\boldsymbol{X}_\ell$ por matrices de pesos $\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V \in \mathbb{R}^{C_\ell \times d}$, donde $d = C_\ell / h_\ell$ es la dimensión por cabeza y h_ℓ el número de cabezas (de modo que la concatenación de las h_ℓ cabezas vuelve a tener dimensión C_ℓ).

En esta arquitectura se definen dos dominios ortogonales de atención mediante franjas (*stripes*) horizontales y verticales con ancho $s_w^{(\ell)}$. La mitad de las cabezas atiende dentro de franjas horizontales y la otra mitad dentro de franjas verticales; así se capta contexto direccional amplio con bajo costo. Tras calcular la atención en cada dominio y cabeza, las salidas se concatenan (recuperando dimensión C_ℓ) y se proyectan para fusionar la información de ambas orientaciones.

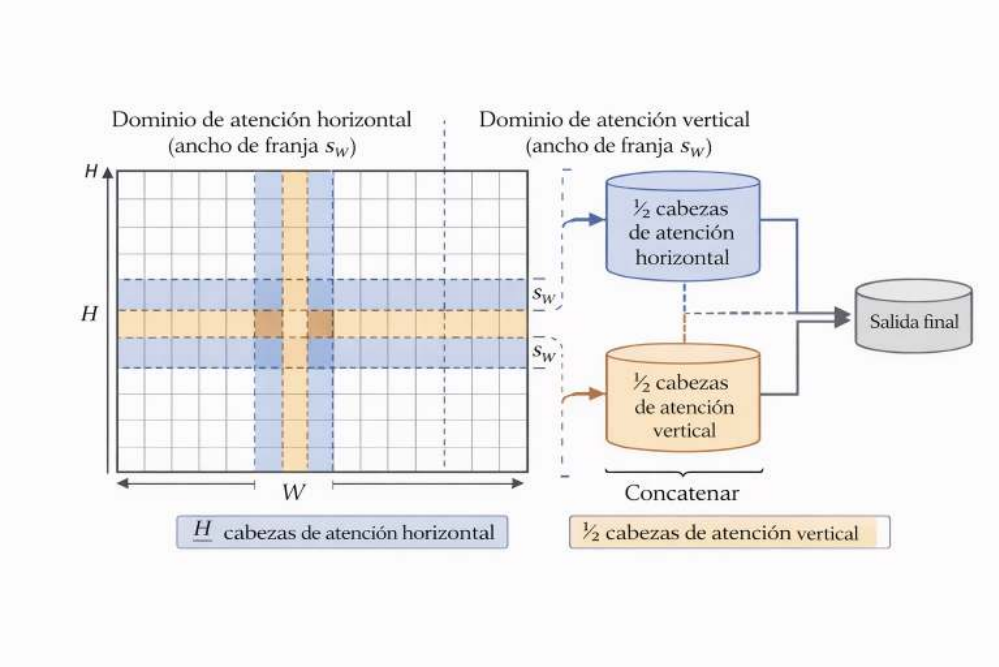


Figura 2.12: Ilustración del mecanismo de autoatención en ventana con forma de cruz. La mitad de las cabezas atiende dentro de franjas horizontales y la otra mitad dentro de franjas verticales de ancho $s_w^{(\ell)}$. Las salidas se concatenan y proyectan para producir la representación final. Adaptada de Dong et al. (2022).

La Figura2.12 ilustra geométricamente esta operación. Para un píxel dado, la atención no se calcula sobre toda la cuadrícula $H_\ell \times W_\ell$, sino únicamente sobre posiciones contenidas en una franja horizontal o vertical de ancho $s_w^{(\ell)}$. La intersección de ambos dominios genera una región con forma de cruz, lo que permite ampliar el campo receptivo sin incurrir en el costo cuadrático de la atención global.

2.7.3.2. Arquitectura jerárquica por etapas.

La red inicia con un *token embedding* convolucional 7×7 con *stride* 4 que lleva la entrada $H \times W$ a $\frac{H}{4} \times \frac{W}{4}$ y C_1 canales. Lo cual convierte la imagen en tokens iniciales con contexto local y reduce

la resolución para contener el coste computacional.

Entre etapas, proyecciones 3×3 con *stride* 2 reducen resolución y suelen *doblar* los canales ($C_{\ell+1} \approx 2C_\ell$). Al disminuir posiciones espaciales, aumentar canales permite representar rasgos más abstractos sin elevar el coste total. (Por ejemplo para 512×512: $128^2 \to 64^2 \to 32^2 \to 16^2$.)

Cada etapa apila N_ℓ bloques con esquema *pre-LN*:

$$\mathbf{U} = \mathbf{X} + \mathrm{CSWinAttn}\big(\mathrm{LN}(\mathbf{X})\big), \tag{2.4}$$

$$\mathbf{Y} = \mathbf{U} + \mathrm{MLP}\big(\mathrm{LN}(\mathbf{U})\big), \tag{2.5}$$

En cada bloque, la normalización previa estabiliza el entrenamiento. La atención del CSWin mezcla información no local y la conexión residual conserva señales. El MLP (dos capas, expansión $4\times$ y proyección a C_ℓ) recombina canales y añade no linealidad.

Donde el MLP es de dos capas con expansión $4\times$ y proyección de retorno a C_ℓ. El *backbone* expone cuatro salidas jerárquicas en resoluciones $\{1/4,\ 1/8,\ 1/16,\ 1/32\}$ de la entrada:

$$\{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4\} \ \text{ con resoluciones } \left\{\tfrac{1}{4},\ \tfrac{1}{8},\ \tfrac{1}{16},\ \tfrac{1}{32}\right\} \text{ de la entrada.} \tag{2.6}$$

Estas salidas multiescala proveen mapas a diferentes niveles de detalle y contexto, ideales para cuellos como FPN o ASPP antes de la cabeza de clasificación.

2.7.3.3. Codificación posicional con realce local.

Para incorporar posición sin fijar una resolución concreta, se introduce *LePE (Locally-enhanced Positional Encoding)* como un ramo paralelo que actúa sobre $\mathbf{V}$ por cabeza (no modifica el puntaje $\mathbf{QK}^\top$). LePE añade un sesgo local que refuerza vecindarios cercanos y preserva la compatibilidad con tamaños de entrada variables, aportando estabilidad cuando se recorta, se hace *padding* o se aplica TTA en inferencia.

2.7.3.4. Complejidad y elección de s_w.

El coste por bloque de la atención ventana en forma de cruz puede escribirse como:

$$\Omega(\mathrm{CSWin}) = H_\ell W_\ell C_\ell \left(4C_\ell + s_w^{(\ell)}\big(H_\ell + W_\ell\big)\right). \tag{2.7}$$

De ese modo existe la necesidad de incrementar $s_w^{(\ell)}$ con la profundidad: pequeño en etapas tempranas (muchos tokens) y mayor en etapas profundas (pocos tokens). El ajuste de referencia comúnmente utilizado es:

$$s_w^{(1..4)} = \big[1,\ 2,\ 7,\ 7\big]. \tag{2.8}$$

Notación utilizada en esta sección: H, W: alto y ancho de la imagen de entrada; H_ℓ, W_ℓ: resolución en la etapa ℓ; C_ℓ: canales (dimensión de *embedding*); N_ℓ: número de bloques por etapa;

h_ℓ: total de cabezas; $d = C_\ell/h_\ell$: dimensión por cabeza; $s_w^{(\ell)}$: ancho de franja por etapa; $\mathbf{M}_\mathcal{D}$: máscara para $\mathcal{D} \in \{\mathrm{H}, \mathrm{V}\}$; $\mathbf{Q}, \mathbf{K}, \mathbf{V}$: consultas, claves y valores por cabeza.

2.8. Estado del arte.

Además de los avances generales en arquitecturas basadas CNN y Transformer, existe un área de trabajo específico orientada al estudio de imágenes de OCT mediante Inteligencia Artificial.

Un avance relevante en la adaptación de Transformers al dominio visual fue propuesto por Liu et al., 2021, quienes introdujeron el Swin Transformer, la cual también es una arquitectura jerárquica basada en atención local mediante ventanas desplazadas. A diferencia del Vision Transformer, Swin restringe el cálculo de auto-atención a ventanas locales no superpuestas, reduciendo la complejidad computacional a un comportamiento lineal con respecto al tamaño de la imagen. El modelo construye representaciones jerárquicas mediante un esquema de *patch merging*, que reduce progresivamente la resolución espacial mientras incrementa la dimensionalidad de los canales, de forma análoga a las redes convolucionales profundas. Este diseño permite que Swin Transformer funcione como un *backbone* general para tareas de clasificación y predicción densa, tales como detección de objetos y segmentación semántica.

Kermany, Goldbaum et al., 2018 constituyen uno de los trabajos fundacionales en la aplicación de aprendizaje profundo a la clasificación automática de patologías retinianas mediante imágenes de Tomografía de Coherencia Óptica (OCT). En este estudio se introduce el conjunto de datos OCT2017 (Kermany, Zhang & Goldbaum, 2018), compuesto por más de 100,000 cortes B etiquetados en cuatro categorías: CNV, DME, DRUSEN y NORMAL, con particiones independientes de entrenamiento y prueba.

Para la tarea de clasificación multiclase, los autores emplean una arquitectura Inception-V3 preentrenada en ImageNet, adaptando la capa final mediante transferencia de aprendizaje. La red profunda funciona como extractor jerárquico de características, aprendiendo representaciones discriminativas directamente a partir de los patrones estructurales presentes en las imágenes OCT.

En el escenario de cuatro clases, el modelo reporta una exactitud global de 96.53 %, junto con valores elevados de sensibilidad y especificidad superiores al 95 %. El estudio demuestra que los modelos de aprendizaje profundo pueden alcanzar un desempeño comparable al de especialistas humanos en la identificación de patologías maculares a partir de imágenes OCT.

Este trabajo marcó un punto de inflexión en la investigación en oftalmología computacional, estableciendo un protocolo experimental ampliamente adoptado y evidenciando la viabilidad del uso de redes profundas para diagnóstico asistido por computadora en imágenes retinianas.

Posterior al trabajo de Kermany, Goldbaum et al., 2018, diversos estudios han continuado explorando el uso de arquitecturas convolucionales profundas, con el objetivo de optimizar el rendimiento en clasificación multiclase mediante estrategias de ajuste fino y optimización avanzada.

En este contexto, Hassan et al., 2023 propusieron un modelo denominado EOCT, basado en una arquitectura ResNet-50 modificada combinada con un clasificador Random Forest y un esquema de optimización dual (SGD y Adam). El modelo fue evaluado sobre el conjunto OCT2017 para la

clasificación en cuatro categorías (CNV, DME, DRUSEN y NORMAL), alcanzando una exactitud de 97.56 % utilizando ResNet-50 con optimizador Adam.

En términos de desempeño balanceado entre clases, el modelo reporta un F1-score de 0.9688, evidenciando una adecuada capacidad discriminativa en las distintas categorías patológicas. Estos resultados confirman que las arquitecturas convolucionales profundas, cuando son adecuadamente optimizadas, continúan ofreciendo un alto rendimiento en tareas de clasificación de imágenes OCT.

Sotoudeh-Paima et al., 2022 proponen una arquitectura multiescala para la clasificación automatizada de patologías maculares a partir de imágenes OCT, basada en una red convolucional profunda con integración tipo Feature Pyramid Network (FPN). El objetivo principal del estudio es explotar explícitamente la información jerárquica presente en las representaciones intermedias de una CNN mediante la fusión de mapas de características a distintas resoluciones espaciales a través de conexiones laterales.

En el caso del conjunto Sotoudeh-Paima et al., 2023, compuesto por más de 16,000 cortes B provenientes de 441 pacientes y etiquetados en tres categorías (NORMAL, DRUSEN y CNV), los autores emplean validación cruzada a nivel de paciente con cinco particiones, junto con estrategias de aumento de datos para mejorar la generalización del modelo. Asimismo, incorporan una función de pérdida ponderada con el fin de compensar el desbalance entre clases.

La configuración con mejor desempeño corresponde a un backbone VGG16 combinado con FPN, alcanzando una exactitud de 92.0 %. Otras configuraciones evaluadas incluyen FPN + DenseNet121 (90.9 %), FPN + ResNet50 (90.1 %) y FPN + EfficientNetB0 (87.8 %). El estudio reporta principalmente la métrica de exactitud para este conjunto en la comparación de arquitecturas.

Los resultados evidencian que la integración multiescala mediante FPN mejora el desempeño respecto al entrenamiento de las arquitecturas base sin fusión piramidal, consolidando la relevancia de estrategias multiescala en la clasificación automatizada de DMAE a partir de imágenes OCT.

Yusufoğlu et al., 2024 proponen MSA-Net (Multi-Scale Attention Network), una arquitectura diseñada específicamente para la clasificación de imágenes OCT en el conjunto de Sotoudeh-Paima et al., 2023. El modelo utiliza EfficientNetB0 como *backbone* principal e incorpora un mecanismo de atención espacial junto con una estrategia de integración multiescala, permitiendo capturar información contextual a diferentes resoluciones espaciales y reforzar las regiones relevantes para la toma de decisión. La arquitectura combina mapas de características intermedios mediante módulos de atención que ponderan dinámicamente la relevancia espacial de las estructuras retinianas, con el objetivo de mejorar la discriminación entre patrones morfológicos sutiles característicos de DRUSEN y CNV.

Siendo el mismo conjunto, compuesto por imágenes etiquetadas en tres categorías (NORMAL, DRUSEN y CNV), los autores emplean una versión reducida del conjunto original, seleccionando 12,649 imágenes provenientes de 441 pacientes. La división experimental se realiza bajo un esquema 9:1, es decir, 90 % de las imágenes se destinan al entrenamiento y 10 % a prueba, sin reportarse un conjunto de validación independiente para la selección del modelo ni un esquema de validación cruzada a nivel de paciente.

En esta configuración experimental, MSA-Net alcanza una exactitud de 98.1 %, una sensibilidad de 97.9 %, una especificidad de 98.0 % y un F1-score de 98.0 % en el conjunto de prueba. El estudio reporta principalmente métricas agregadas tradicionales de clasificación, sin incluir indicadores adicionales como AUC, Macro-F1, análisis de curvas ROC por clase o métricas que evalúen

explícitamente el desempeño balanceado ante posibles desbalances interclase.

Si bien los resultados evidencian un desempeño elevado en la clasificación de DMAE sobre el conjunto mencionado anteriormente, la ausencia de una validación independiente o cruzada limita la evaluación de la estabilidad del modelo frente a variaciones en la partición de datos. Asimismo, el uso de una única división 90/10 puede introducir sesgos dependientes del subconjunto seleccionado, dificultando la comparación directa bajo protocolos experimentales más estrictos. No obstante, el trabajo destaca la relevancia de integrar mecanismos de atención espacial con estrategias multiescala para mejorar la discriminación estructural en imágenes OCT, consolidando la tendencia hacia arquitecturas que combinan eficiencia convolucional con refinamiento atencional.

Otros trabajos recientes en OCT exploran Transformers o híbridos CNN–Transformer para tareas de segmentación y clasificación, pero en su mayoría se enfocan en retinopatía diabética o segmentación de capas retinianas Arsalan et al., 2022; Shi et al., 2023, y no emplean el conjunto de datos OCT etiquetado específicamente para DMAE utilizado en este trabajo.

En este contexto, este trabajo sitúa en la intersección entre: (i) el uso de un backbone Transformer jerárquico de última generación (CSWin-Transformer) adaptado a OCT en escala de grises, y (ii) la integración con cuellos multiescala (FPN y ASPP) para explotar explícitamente la información jerárquica de las imágenes, evaluando su efecto sobre métricas macro (macro-F1 y AUC macro) en comparación con los enfoques CNN clásicos reportados en el estado del arte.

2.9. Arquitectura Híbrida CSWin-Transformer con FPN.

En este punto se plantea una arquitectura de clasificación que combina el CSWin-Transformer como *backbone* jerárquico con un cuello FPN para integrar de manera multi-escalar antes de la cabeza final. De modo que se tiene la siguiente organización:

Estructura

- **Entrada y proyección inicial.**
- **Backbone CSWin-Transformer.**
- **Cuello FPN.**
- **Agregación y cabeza.**

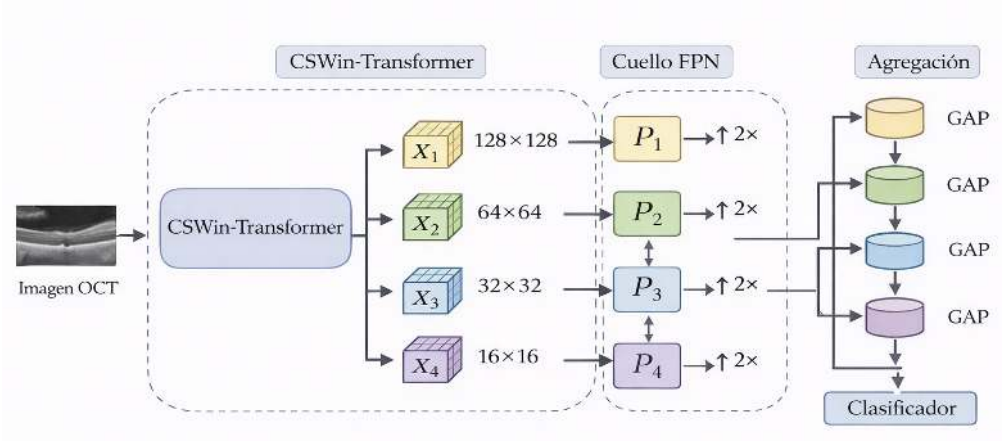


Figura 2.13: Diagrama arquitectura CSWin-Transformer+FPN.

La Figura 2.13 muestra la arquitectura híbrida propuesta. El CSWin-Transformer produce cuatro mapas de características multiescala $\{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4\}$, correspondientes a resoluciones espaciales $\{128 \times 128,\ 64 \times 64,\ 32 \times 32,\ 16 \times 16\}$ para una entrada de 512×512.

Estos mapas son enviados al cuello FPN, donde cada nivel se proyecta mediante convoluciones 1×1 para unificar la dimensión de canales y generar los mapas piramidales $\{\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3, \mathbf{P}_4\}$. Posteriormente, se realiza una integración de arriba hacia abajo (*top–down pathway*) mediante operaciones de *upsampling* y sumas laterales, permitiendo combinar semántica profunda (niveles altos) con detalle espacial fino (niveles bajos).

En la etapa de agregación final, se aplica *Global Average Pooling* (GAP) a cada mapa $\mathbf{P}\ell$. Esta operación promedia los valores espaciales de cada canal y transforma un tensor de dimensiones $H\ell \times W_\ell \times C$ en un vector de dimensión C, conservando la activación media global de cada filtro. Su uso permite obtener una representación compacta de cada nivel de la pirámide y, al mismo tiempo, reducir el número de parámetros respecto a estrategias basadas en el aplanamiento completo de los mapas de características, lo que contribuye a disminuir el riesgo de sobreajuste.

Finalmente, los vectores resultantes se concatenan para formar una representación holística multiescala, la cual se introduce en la cabeza de clasificación completamente conectada encargada de predecir las clases clínicas correspondientes. De esta manera, la decisión final del modelo se apoya en información integrada de distintos niveles de resolución, lo que favorece una clasificación más robusta ante variaciones anatómicas y morfológicas presentes en las imágenes OCT.

El diseño de esta arquitectura resulta adecuado para imágenes OCT por varias razones. En primer lugar, la FPN introduce una integración multiescala explícita, al combinar la semántica profunda de los niveles superiores del backbone con el detalle espacial fino preservado en los niveles inferiores, lo cual es especialmente útil para reconocer patrones sutiles como pequeñas drusas o variaciones en los bordes del epitelio pigmentario retiniano. En segundo lugar, se trata de una estrategia computacionalmente eficiente, ya que la mayor parte del costo se concentra en el backbone y el cuello FPN añade únicamente proyecciones laterales y operaciones de *upsampling* relativamente ligeras. Finalmente, la agregación global de los mapas piramidales permite construir una representación holística menos sensible a variaciones de escala, geometría y distribución espacial de los hallazgos retinianos, lo que favorece una clasificación más robusta en presencia de variabilidad anatómica entre pacientes.

2.10. Arquitectura Híbrida CSWin-Transformer con ASPP.

En segundo punto, se plantea una arquitectura de clasificación que combina el CSWin-Transformer como *backbone* jerárquico con un módulo Atrous Spatial Pyramid Pooling (ASPP) como cuello para la integración multiescala basada en dilataciones antes de la cabeza final. De modo que se tiene la siguiente organización:

Estructura

- **Entrada y proyección inicial.**
- **Backbone CSWin-Transformer.**
- **Módulo ASPP.**
- **Agregación y cabeza.**

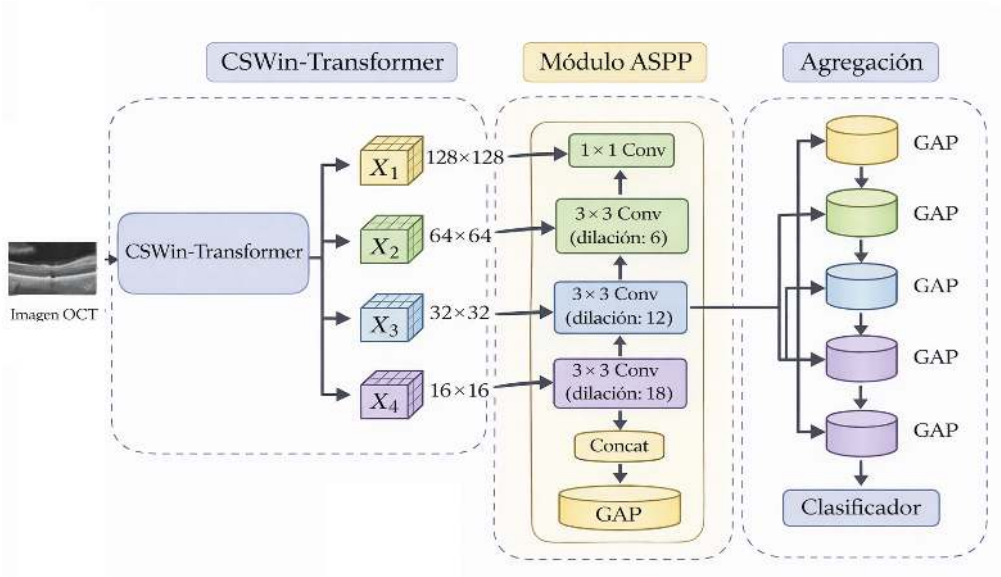


Figura 2.14: Diagrama arquitectura CSWin-Transformer+ASPP.

La Figura 2.14 muestra la arquitectura híbrida propuesta. El CSWin-Transformer produce cuatro mapas de características multiescala $\{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4\}$, correspondientes a resoluciones espaciales $\{128 \times 128,\ 64 \times 64,\ 32 \times 32,\ 16 \times 16\}$ para una entrada de 512×512.

A diferencia del enfoque con FPN, en esta configuración el módulo ASPP se aplica sobre el mapa de mayor nivel semántico (típicamente $\mathbf{X}_4$), el cual contiene representaciones profundas y altamente abstractas de la estructura retiniana.

El ASPP implementa múltiples convoluciones paralelas con diferentes tasas de dilatación (*atrous rates*), permitiendo capturar contexto a distintos campos receptivos sin reducir la resolución espacial. Formalmente, si $\mathbf{X}_4 \in \mathbb{R}^{H \times W \times C}$, el ASPP genera un conjunto de mapas:

$$\{\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{A}_4\}$$

donde cada $\mathbf{A}_k$ corresponde a una convolución con distinta dilatación. Adicionalmente, se incorpora una rama de *Global Average Pooling* (GAP) que captura contexto global completo. Posteriormente, todas las ramas se concatenan y se proyectan mediante una convolución 1×1 para fusionar la información multiescala en un solo tensor enriquecido.

En la etapa de agregación final, se aplica *Global Average Pooling* sobre el mapa resultante del ASPP, transformando el tensor espacial en un vector compacto de dimensión C', el cual resume la información contextual integrada a múltiples escalas.

Finalmente, esta representación global se introduce en la cabeza de clasificación completamente conectada encargada de predecir las clases clínicas correspondientes.

¿Por qué diseñar una arquitectura de este estilo para OCT?:

(i) *Contexto multiescala implícito:* las convoluciones dilatadas permiten capturar simultáneamente estructuras pequeñas (drusas incipientes) y patrones más extensos (desprendimientos, neovascularización) sin perder resolución espacial.

(ii) *Preservación de detalle estructural:* al no requerir un proceso de fusión jerárquica explícita como FPN, el ASPP mantiene intacta la representación profunda del último nivel, refinándola mediante distintos campos receptivos.

(iii) *Mayor discriminación contextual:* la combinación de dilataciones y contexto global favorece la separación entre clases con patrones morfológicos similares pero distribuciones espaciales distintas, lo cual resulta especialmente relevante en imágenes OCT donde la variación anatómica es considerable.

La incorporación de ASPP se justifica en debido a que las manifestaciones estructurales de interés en imágenes OCT no se presentan en una única escala espacial. En particular, las drusas suelen observarse como elevaciones focales relativamente pequeñas del epitelio pigmentario retiniano, mientras que las alteraciones asociadas con CNV pueden involucrar patrones más extensos, irregulares o dependientes del contexto anatómico circundante. En este escenario, ASPP permite explorar simultáneamente distintos campos receptivos sobre una representación profunda de alto contenido semántico, mediante ramas paralelas con diferentes tasas de dilatación y una rama adicional de contexto global. Esta estrategia favorece la integración de información local, intermedia y global sin requerir una reducción adicional de la resolución espacial del mapa de entrada, por lo que resulta adecuada para enriquecer la discriminación entre las categorías NORMAL, DRUSEN y CNV en un problema de clasificación multiclase basado en OCT. Esta variante busca aprovechar la capacidad del backbone Transformer para modelar relaciones espaciales complejas y complementarla con una integración multiescala explícita centrada en la representación profunda final.

2.11. Interpretabilidad visual.

En aplicaciones biomédicas basadas en aprendizaje profundo, no basta con alcanzar un desempeño cuantitativo elevado; también es importante contar con herramientas que permitan analizar, de forma visual, qué regiones de la imagen influyen en la decisión del modelo. Este tipo de análisis resulta especialmente valioso en imágenes médicas, ya que ayuda a determinar si la predicción se apoya en zonas anatómicamente concretas y no en patrones irrelevantes. En este contexto, dos recursos de interpretabilidad ampliamente utilizados son *Grad-CAM* y los mapas de atención.

Gradient-weighted Class Activation Mapping (Grad-CAM) es una técnica de interpretabilidad visual que permite identificar qué regiones espaciales contribuyen con mayor intensidad a la predicción de una clase determinada. Su principio consiste en utilizar los gradientes de la salida asociada a la clase de interés con respecto a mapas de características internos del modelo para construir un mapa de activación discriminativo por clase. De este modo, se obtiene una representación visual que resalta las zonas que tuvieron mayor peso en la decisión final. Una de las principales ventajas de Grad-CAM es que puede aplicarse a distintas arquitecturas basadas en convoluciones sin necesidad de modificar el modelo ni reentrenarlo, lo que ha favorecido su adopción en tareas de análisis visual y en contextos donde la interpretabilidad es relevante.

En el ámbito de la imagen médica, Grad-CAM se utiliza con frecuencia como apoyo para el análisis cualitativo del comportamiento del modelo, ya que permite verificar si la red concentra su respuesta en estructuras anatómicas compatibles con el problema estudiado. Aunque no sustituye una validación clínica ni una evaluación cuantitativa formal, sí aporta una evidencia visual útil para examinar la coherencia de las predicciones y detectar posibles sesgos en el proceso de aprendizaje.

Por su parte, los mapas de atención corresponden a representaciones derivadas de los mecanismos de atención empleados por los modelos tipo Transformer. En términos generales, la atención asigna pesos relativos a distintas posiciones de la entrada para construir una representación contextualizada, de modo que puede analizarse qué regiones o tokens reciben mayor importancia durante el procesamiento. En visión por computadora, este principio se traslada a parches o regiones de la imagen, lo que permite estudiar cómo el modelo distribuye su enfoque entre información local y contextual.

En el caso particular del CSWin-Transformer, esta interpretación resulta especialmente pertinente, ya que su mecanismo de atención organiza la interacción espacial mediante ventanas cruzadas horizontales y verticales. Esta estrategia permite modelar relaciones más amplias que una atención puramente local, preservando al mismo tiempo una complejidad computacional controlada. Por ello, el análisis de mapas de atención en esta arquitectura ofrece una referencia cualitativa útil para observar cómo se propaga la información entre distintas regiones de la OCT y cómo el modelo integra detalles locales con contexto estructural más amplio.

En conjunto, Grad-CAM y los mapas de atención constituyen herramientas complementarias para la interpretación visual del modelo. Mientras Grad-CAM permite resaltar regiones de la imagen asociadas directamente con la clase predicha, los mapas de atención ayudan a examinar cómo se distribuyen internamente las relaciones entre regiones durante la extracción de características. Así, ambas técnicas enriquecen el análisis cualitativo del sistema propuesto y fortalecen la interpretación de sus resultados desde una perspectiva más transparente.

CAPÍTULO 3

Hipótesis.

Un modelo de aprendizaje profundo basado en inteligencia artificial, entrenado con imágenes de Tomografía de Coherencia Óptica preprocesadas, mejorará el desempeño diagnóstico para clasificar casos NORMAL, DRUSAS y CNV (asociados a biomarcadores de DMAE) en comparación con los modelos base descritos en el estado del arte, evaluado en el conjunto de prueba mediante macro-F1 y AUC (one-vs-rest), mostrando un incremento mínimo en métricas propias de los algoritmos de inteligencia artificial.

CAPÍTULO 4

Objetivos.

4.1. Objetivo general.

Desarrollar y evaluar un algoritmo de aprendizaje profundo para el análisis de imágenes de Tomografía de Coherencia Óptica (OCT), con el fin de mejorar la detección temprana y la clasificación de casos NORMAL, DRUSEN y CNV asociados a la Degeneración Macular Asociada a la Edad (DMAE), mediante el preprocesamiento estandarizado de las imágenes, el entrenamiento y validación del modelo, y su comparación con métodos convencionales y modelos base del estado del arte usando métricas como macro-F1 y AUC(one-vs-rest).

4.2. Objetivos específicos.

- Diseñar una arquitectura de aprendizaje profundo para identificar patrones característicos de la DMAE en OCT, mediante la selección de una red adecuada y módulos de extracción de características multiescala.
- Obtener una base de datos pública de imágenes OCT etiquetadas para entrenar y evaluar el modelo propuesto, mediante su descarga, organización por clases (NORMAL, DRUSAS, CNV) y verificación de integridad y consistencia de etiquetas.
- Preprocesar la base de datos obtenida para estandarizar la entrada del modelo y reducir variaciones no clínicas, mediante recorte de región de interés (ROI), normalización, redimensionamiento y partición en conjuntos de entrenamiento, validación y prueba.
- Implementar el modelo de aprendizaje profundo para ejecutar entrenamiento e inferencia de forma reproducible, mediante el desarrollo del pipeline en el framework seleccionado, con control de semillas y registro de experimentos.
- Entrenar el modelo para clasificar automáticamente imágenes OCT en NORMAL, DRUSAS y CNV y analizar zonas relevantes, mediante estrategias de entrenamiento (fine-tuning, regu-

larización y aumento de datos) y, si aplica, técnicas de interpretabilidad (p. ej. Grad-CAM o mapas de atención).

- Evaluar la eficacia del método propuesto para determinar su superioridad frente a enfoques convencionales y modelos base, mediante macro-F1, AUC(one-vs-rest), matriz de confusión y pruebas estadísticas.
- Ajustar y optimizar el modelo para mejorar su desempeño y capacidad de generalización, mediante ajuste de hiperparámetros (tasa de aprendizaje, regularización, tamaño de entrada, número de épocas) y análisis de errores por clase.

CAPÍTULO 5

Metodología.

Este capítulo describe el procedimiento seguido para construir, entrenar y evaluar un sistema de clasificación multiclase de Degeneración Macular Asociada a la Edad (DMAE) en imágenes de Tomografía de Coherencia Óptica (OCT), utilizando arquitecturas híbridas basadas en CSWin-Transformer con módulos de agregación multiescala FPN y ASPP.

5.1. Base de datos.

Se utilizó la base de datos *Labeled Retinal Optical Coherence Tomography Dataset for Classification of Normal, Drusen, and CNV Cases*, disponible en Mendeley Data (Sotoudeh-Paima et al., 2023). El conjunto contiene 16,822 imágenes B-scan de OCT provenientes de 441 pacientes atendidos en el Noor Eye Hospital. Los casos incluidos se distribuyen en tres categorías clínicas: 120 pacientes normales, 160 pacientes con drusas y 161 pacientes con neovascularización coroidea (CNV).

Las imágenes se encuentran clasificadas en tres categorías diagnósticas: *NORMAL*, *DRUSEN* y *CNV*. Dado que las etiquetas fueron asignadas con base en criterios clínicos y oftalmológicos especializados, el conjunto constituye una fuente adecuada para el entrenamiento y evaluación de modelos de clasificación automática en OCT.

Los datos cuentan con licencia Creative Commons Attribution 4.0 (CC BY 4.0), lo que permite su uso abierto siempre que se otorgue el crédito correspondiente. Esta base se asocia al trabajo de Sotoudeh-Paima et al. (2022), donde se empleó para el desarrollo de modelos convolucionales multiescala.

5.1.1. Entorno de trabajo.

Hardware: se utilizó una GPU NVIDIA A100.

Plataforma de desarrollo: Google Colaboratory (Google Colab), en modalidad Colab Pro, se empleó como entorno principal para el desarrollo, entrenamiento y evaluación de los modelos.

5.2. Obtención del conjunto de datos.

El conjunto se descargó desde Mendeley Data (Sotoudeh-Paima et al., 2023) (fecha de acceso: 01 de septiembre de 2025). La descarga se realizó mediante la interfaz web, y los archivos fueron copiados a Google Drive para su uso posterior en Google Colab. La estructura final se organizó dentro de la ruta del proyecto, preservando los nombres originales de archivo con el fin de mantener la trazabilidad y reproducibilidad experimental.

5.2.1. Filtrado intra-paciente basado en metadatos del CSV.

El conjunto original contiene múltiples B-scans por paciente y, de acuerdo con el archivo CSV de metadatos proporcionado junto con la base, cada B-scan posee una etiqueta individual que puede diferir dentro de un mismo volumen (Sotoudeh-Paima et al., 2022).

Se observó que en pacientes asignados a una categoría clínica dominante, por ejemplo CNV, podían existir cortes individuales etiquetados como *NORMAL* o *DRUSEN*, correspondientes a regiones del volumen donde la lesión no era claramente visible. Dado que el presente trabajo realiza clasificación a nivel de imagen individual (B-scan), la presencia de cortes ambiguos o inconsistentes introduce ruido supervisado y reduce la fuerza de la señal patológica.

Por ello, se implementó un procedimiento determinístico basado en el CSV para:

- Agrupar las imágenes por identificador de paciente.
- Conservar únicamente los B-scans cuya etiqueta individual coincidía con la condición clínica dominante del paciente.
- Eliminar cortes inconsistentes o con evidencia mínima de la patología.

Este filtrado redujo el conjunto de 16,822 a 12,649 imágenes, manteniendo los 441 pacientes originales. El procedimiento se realizó antes de cualquier partición experimental y no utilizó información derivada del desempeño del modelo, garantizando ausencia de fuga de información.

5.3. División del conjunto y protocolo experimental.

Una vez realizado el filtrado intra-paciente, se definió el protocolo experimental principal del estudio. En lugar de emplear una única partición fija de entrenamiento, validación y prueba, se adoptó un esquema compuesto por un **conjunto de prueba externo fijo** y una **validación cruzada estratificada y agrupada de 5 folds** sobre el conjunto restante, con el fin de garantizar una comparación justa, todos los experimentos realizados en este trabajo utilizaron el mismo conjunto de prueba externo y la misma partición interna, de modo de cualquier diferencia puede ser atribuida a la arquitectura y no a varaciones en la partición de datos.

5.3.1. Construcción del conjunto de prueba externo.

El primer paso consistió en separar un 20 % de los datos como conjunto de prueba externo, el cual permaneció fijo durante todos los experimentos. Esta separación se realizó a nivel de paciente, con el fin de evitar que imágenes de un mismo individuo aparecieran simultáneamente en entrenamiento y prueba.

Para ello, se construyó un identificador de grupo por paciente, utilizado como unidad de partición. De manera complementaria, se preservó la distribución de clases clínicas durante la separación, de modo que el conjunto de prueba mantuviera representatividad diagnóstica comparable al conjunto global filtrado.

Como resultado, se obtuvo:

- **Conjunto de desarrollo** (*train outer base*): 10,128 imágenes.
- **Conjunto de prueba externo fijo**: 2,521 imágenes.

Tabla 5.1: Partición externa del conjunto de datos tras el filtrado intra-paciente.

Subconjunto	**Número de imágenes**
Conjunto de desarrollo (80 %)	10,128
Conjunto de prueba externo (20 %)	2,521
Total	**12,649**

5.3.2. Validación cruzada estratificada y agrupada.

Sobre el conjunto de desarrollo se aplicó una validación cruzada de 5 folds mediante `StratifiedGroupKFold`. Esta estrategia permitió cumplir simultáneamente dos objetivos:

- **Estratificación**: conservar la proporción de clases en cada fold.
- **Agrupación**: impedir que imágenes del mismo paciente aparecieran al mismo tiempo en entrenamiento y validación.

En cada iteración, cuatro folds se utilizaron para entrenamiento y uno para validación. De esta manera, se entrenaron cinco modelos independientes por arquitectura, cada uno con su correspondiente subconjunto de entrenamiento y validación, mientras que el conjunto de prueba externo permaneció inalterado para la evaluación final.

Este diseño experimental ofrece una estimación más robusta del desempeño que una única partición, al reducir la dependencia de una sola división y controlar de forma explícita la fuga de información entre pacientes.

5.4. Construcción del dataset base único.

Una vez definido el protocolo experimental, se construyó un **dataset base único** a partir del conjunto de desarrollo. El objetivo fue materializar cada imagen una sola vez tras el preprocesamiento inicial, evitando duplicaciones innecesarias entre folds y reduciendo el costo de almacenamiento y entrada/salida.

En primer lugar, se generó un índice base del conjunto de desarrollo, en el cual cada muestra quedó asociada a su ruta, clase e identificador de grupo. Posteriormente, a partir de este índice base, se construyeron los índices específicos de cada fold, indicando qué muestras pertenecían al subconjunto de entrenamiento y cuáles al de validación en cada iteración.

Este procedimiento permitió:

- evitar la redundancia de imágenes entre folds,
- mantener trazabilidad entre la imagen original y su representación preprocesada,
- construir subconjuntos dinámicos a partir de índices en lugar de copias físicas,
- y verificar explícitamente la ausencia de fuga de pacientes entre entrenamiento y validación.

Así, los folds no se materializaron como conjuntos de imágenes independientes, sino como particiones lógicas definidas por archivos CSV e índices asociados al dataset base único.

5.5. Preprocesamiento de la base de datos.

El preprocesamiento se diseñó con dos objetivos principales: estandarizar geométricamente las imágenes OCT y resaltar la región anatómica relevante para la clasificación. El pipeline se aplicó de manera determinística a todas las muestras del dataset base, mientras que la aumentación de datos se reservó exclusivamente para los subconjuntos de entrenamiento de cada fold.

5.5.1. Estandarización geométrica inicial.

Cada imagen fue convertida a una representación monocanal explícita y reordenada al formato tensorial requerido por PyTorch. Posteriormente:

1. Se escalaron las intensidades a $[0, 1]$ cuando las imágenes provenían de `uint8`; en caso contrario, se convirtieron a `float32` y se limitaron al mismo rango.
2. Se aplicó **padding simétrico** con modo `reflect` hasta obtener una forma estándar de $[1, 512, 768]$ (canal $\times$ alto $\times$ ancho), centrando la imagen sin introducir deformaciones geométricas por *resize*.
3. Se definió un mapeo fijo de etiquetas:

$$\{\texttt{NORMAL} \to 0, \quad \texttt{DRUSEN} \to 1, \quad \texttt{CNV} \to 2\}.$$

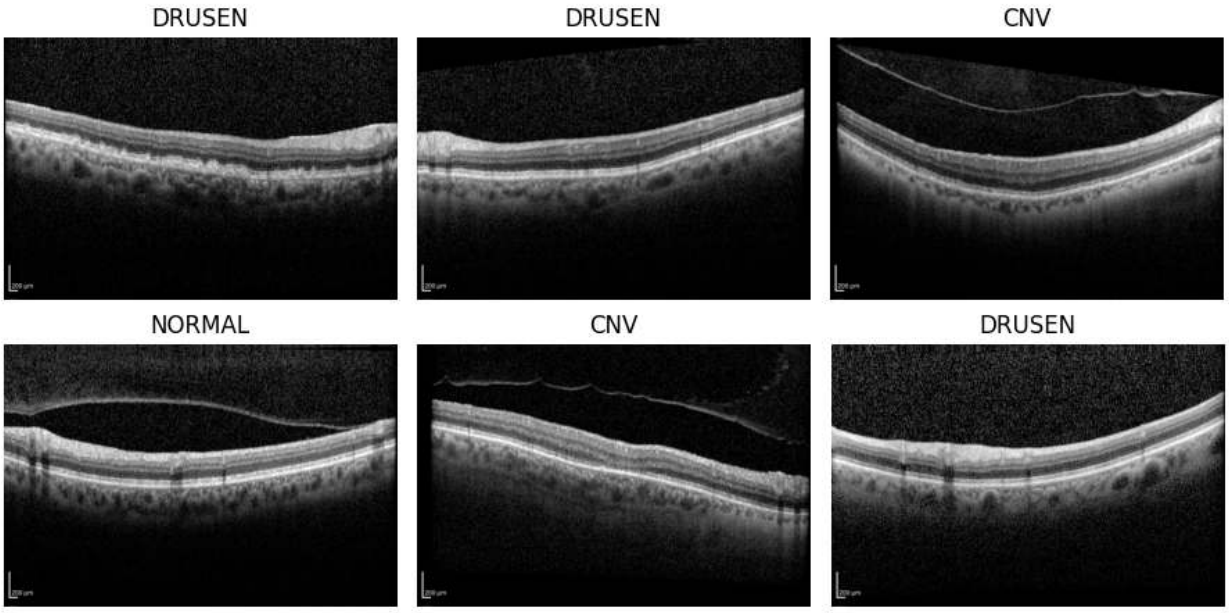


Figura 5.1: Visualización aleatoria de la estandarización geométrica aplicada a algunas imágenes del conjunto.

5.5.2. Materialización del dataset base en *shards*.

Para optimizar el acceso a disco y el uso de memoria, el dataset base se materializó en *shards* `.npy` mediante un recorrido *streaming*. Se utilizaron los siguientes parámetros:

- `shard_size` = 128
- `batch_size` = 16
- `num_workers` = 0
- `to_uint8` = `True`

Cada lote fue cuantizado a `uint8` tras asegurar que las intensidades permanecieran en $[0, 1]$. Esta decisión redujo el tamaño en disco y aceleró la lectura posterior sin comprometer la semántica estructural relevante de las imágenes, dado que las normalizaciones de intensidad más finas se aplicaron posteriormente.

Se generaron pares de archivos `X` e `y` por *shard*, así como archivos auxiliares con metadatos de forma, tipo de dato y mapeo de clases.

5.5.3. Lectura y verificación del dataset materializado.

Se implementó un mecanismo de lectura con `mmap` para evitar cargar los *shards* completos en memoria RAM. Como control de calidad se realizaron las siguientes comprobaciones:

1. recuperación aleatoria de lotes para visualización,

2. inspección visual de B-scans con sus etiquetas,

3. cálculo de estadísticas por imagen para detectar rangos atípicos.

El uso de `reflect padding` evitó la introducción de bordes artificiales negros que pudieran sesgar el aprendizaje de filtros o mecanismos de atención.

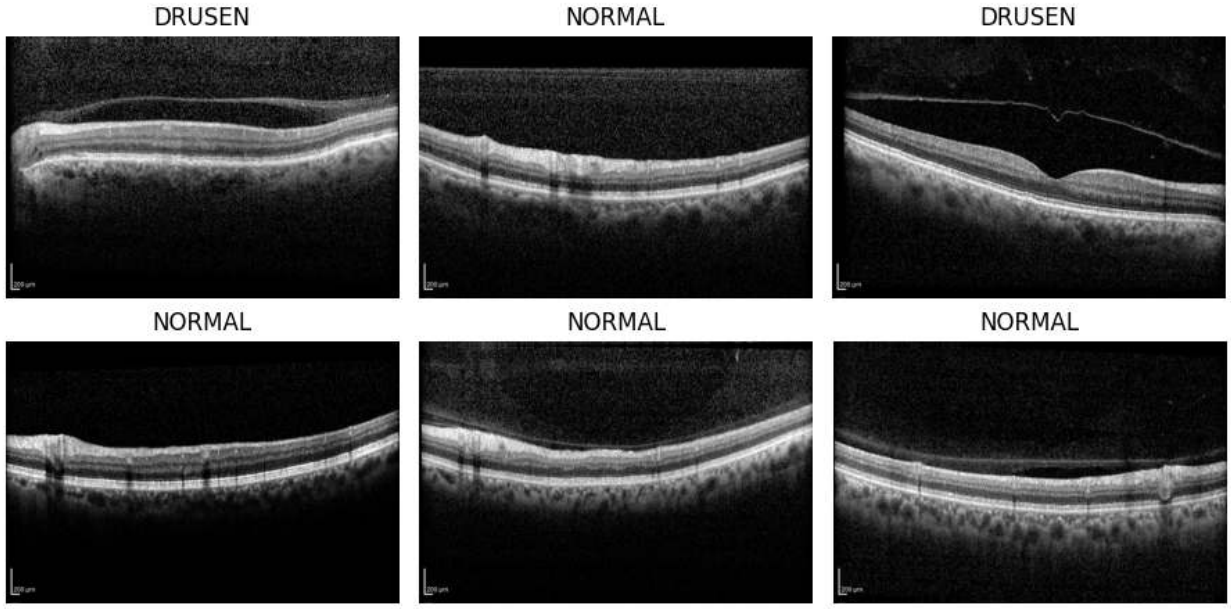


Figura 5.2: Visualización aleatoria de muestras recuperadas desde los *shards* materializados.

5.5.4. Región de interés y estandarización a 512×512.

A partir de la imagen normalizada en $[0, 1]$, se estimó una región de interés (ROI) alrededor de la retina mediante un umbral de intensidad $I > 10^{-6}$. Con la máscara binaria resultante se obtuvo la caja delimitadora mínima del contenido útil y se expandió con un margen de 16 píxeles en cada dirección para preservar contexto anatómico.

El recorte obtenido se ajustó posteriormente a un formato cuadrado mediante *letterbox* a **512×512** píxeles, manteniendo la relación de aspecto y centrando el contenido sobre un lienzo uniforme. Además, se generó una máscara rectangular que identifica la región válida del ROI dentro del lienzo; esta máscara se utilizó posteriormente para la normalización de intensidad. De modo que esto permitió concentrar el análisis en la zona anatómicamente importante de la OCT y reducir la influencia de regiones de fondo o áreas sin contenido clínicamente relevante, del mismo modo, el uso de letterbox preservó la relación de aspecto del contenido retinal durante el ajuste a 512×512, evitando deformaciones geométricas artificiales que podrían alterar patrones morfológicos sutiles de interés diagnóstico.

5.5.5. Aumentación geométrica y fotométrica.

Las transformaciones de aumentación se aplicaron exclusivamente al subconjunto de entrenamiento de cada fold, una vez que las imágenes habían sido recortadas al ROI y ajustadas a 512×512. La misma transformación geométrica se aplicó también a la máscara del ROI con el fin de preservar la coherencia entre la imagen y la región válida. A diferencia de un esquema en el que se define de antemano un subconjunto fijo de imágenes aumentadas, en este trabajo la aumentación se realizó de forma dinámica durante la carga de datos, de modo que cada imagen de entrenamiento fue considerada candidata a recibir transformaciones en cada presentación y podía aparecer modificada de manera distinta entre épocas.

Las transformaciones utilizadas fueron:

- **Rotación**: hasta $\pm 5°$, con probabilidad 0.5.
- **Desplazamiento horizontal**: hasta 3 % del ancho, con probabilidad 0.5.
- **Volteo horizontal**: deshabilitado, por sensibilidad a la lateralidad en OCT.
- **Brillo y contraste**: variaciones multiplicativas de ± 10 %, con probabilidad 0.5 cada una.

Dado que la rotación, el desplazamiento horizontal y los ajustes de brillo y contraste se aplicaron de manera independiente con probabilidad $p = 0.5$, la probabilidad de que una imagen recibiera al menos una transformación en una presentación fue de $1-(0.5)^4 = 0.9375$, es decir, aproximadamente 93.75 %. En consecuencia, sólo un 6.25 % de las presentaciones permanecieron sin modificación en una iteración determinada. Las transformaciones afines se realizaron con *padding* por reflexión para evitar artefactos de borde, mientras que los subconjuntos de validación y prueba externa se mantuvieron sin aumentación, con el fin de obtener estimaciones imparciales del desempeño.

Las transformaciones geométricas se mantuvieron en rangos conservadores para simular pequeñas variaciones plausibles de adquisición y posicionamiento, mientras que los ajustes fotométricos permitieron reducir la sensibilidad del modelo a cambios moderados de brillo y contraste entre estudios. La exclusión deliberada del volteo horizontal obedeció a la necesidad de preservar la coherencia anatómica y evitar alteraciones artificiales asociadas con la lateralidad de la imagen OCT.

5.5.6. Normalización Z-Score centrada en el ROI.

Se aplicó una normalización *z-score* por imagen utilizando únicamente los píxeles válidos del ROI definidos por la máscara del *letterbox*. La transformación fue:

$$I^\star = \frac{I - \mu_\Omega(I)}{\sigma_\Omega(I) + 10^{-6}},$$

donde Ω denota la región válida del ROI y 10^{-6} actúa como término de estabilidad numérica. Esta normalización se aplicó en entrenamiento, validación y prueba externa.

La elección de una normalización z-score se justifica mediante su capacidad para estandarizar la distribución de intensidades de cada imagen con respecto a su propia media y desviación estándar,

reduciendo así variaciones globales de brillo y contraste que no necesariamente corresponden a diferencias patológicas, sino a condiciones de adquisición o al rango dinámico de cada estudio. En este trabajo, dicha normalización se aplicó únicamente sobre la región válida de la imagen, con el fin de evitar que el fondo introducido por el ajuste geométrico o por zonas sin información anatómica sesgara los parámetros de normalización.

A diferencia de una normalización *min-max*, que reescala la imagen en función de sus valores extremos y puede ser más sensible a ruido, valores atípicos o regiones de fondo dominantes, la normalización *z-score* permite homogeneizar la distribución interna de intensidades de una manera más estable. Esto resulta particularmente útil en imágenes OCT, donde interesa preservar la relación relativa entre estructuras retinales y reducir la variabilidad entre muestras sin comprimir artificialmente el contraste relevante dentro de la región anatómica analizada. En consecuencia, la normalización *z-score* centrada en el ROI favoreció una entrada más consistente para el modelo y más adecuada para la comparación entre imágenes.

5.5.7. Carga por lotes y comprobaciones.

Los *DataLoaders* se construyeron dinámicamente para cada fold a partir de los índices definidos en los CSV correspondientes. En todos los casos se empleó:

- **batch size = 16**
- **num_workers = 2**
- `pin_memory` activado

El parámetro `shuffle` se habilitó únicamente en entrenamiento. Como control adicional se verificó la ausencia de valores `NaN`/`Inf`, y se revisaron estadísticas por imagen tras la normalización.

5.5.8. Visualización por clase.

Con el objetivo de verificar la coherencia del preprocesamiento y detectar posibles anomalías antes del entrenamiento, se realizó una inspección cualitativa de muestras pertenecientes a los subconjuntos de entrenamiento, validación y prueba externa. Para cada subconjunto se visualizaron ejemplos de las tres clases diagnósticas: *NORMAL*, *DRUSEN* y *CNV*.

Las imágenes se mostraron en escala de grises utilizando una ventana fija sobre intensidades normalizadas por *z-score*, con rango $[-2, 2]$ reescalado a $[0, 1]$ únicamente con fines de visualización. Asimismo, se dibujó el contorno del área válida del ROI en color verde para evidenciar la región útil del contenido anatómico.

5.6. Arquitectura híbrida.

Se estudiaron dos variantes arquitectónicas que comparten el mismo backbone y difieren únicamente en el módulo de agregación multiescala: una variante con **Feature Pyramid Network (FPN)** y otra con **Atrous Spatial Pyramid Pooling (ASPP)**.

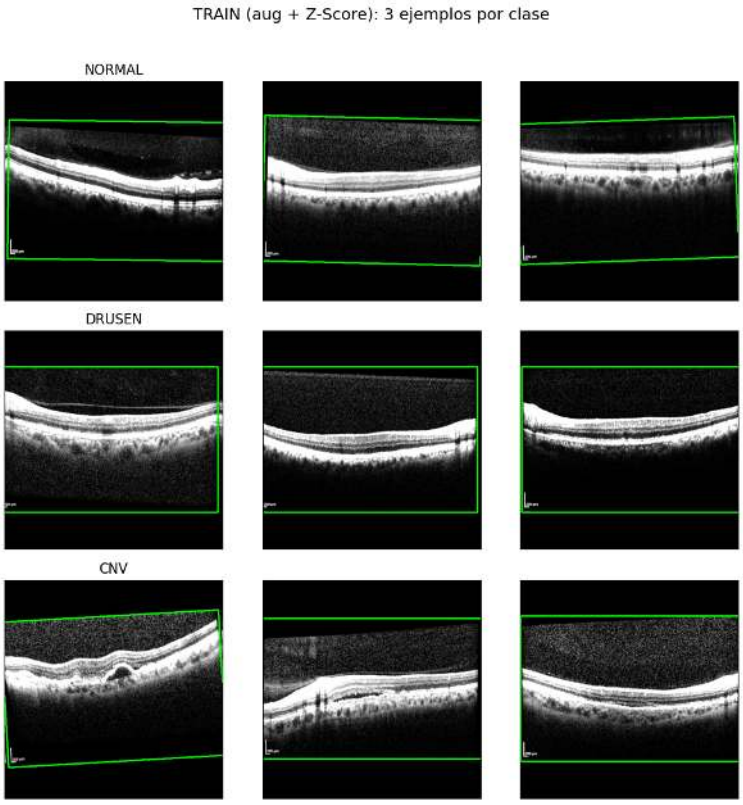


Figura 5.3: Visualización de muestras del subconjunto de entrenamiento con aumentación y normalización *z-score*.

5.6.1. Backbone: CSWin-Transformer.

Se empleó un backbone CSWin-Transformer configurado para imágenes monocanal de 512×512. La instancia utilizada correspondió a la variante *tiny*, con los siguientes hiperparámetros:

$$patch\_size = 4, \quad embed\_dim = 64, \quad depth = [1, 2, 21, 1], \quad num\_heads = [2, 4, 8, 16],$$

$$mlp\_ratio = 4.0, \quad drop\_rate = 0.0, \quad attn\_drop\_rate = 0.0, \quad drop\_path\_rate = 0.2.$$

Para hacer compatible la atención por ventanas con la resolución de entrada, se fijó:

$$split\_size = [1, 2, 8, 8].$$

El backbone se inicializó a partir del checkpoint oficial preentrenado `cswin_tiny_224.pth`. Dado que el modelo original estaba entrenado con imágenes RGB, los pesos de la primera capa se adaptaron a monocanal promediando los filtros a través de los tres canales de entrada. Los parámetros incompatibles en la primera capa y la cabeza original fueron excluidos de la carga, manteniéndose el resto de pesos preentrenados de manera no estricta.

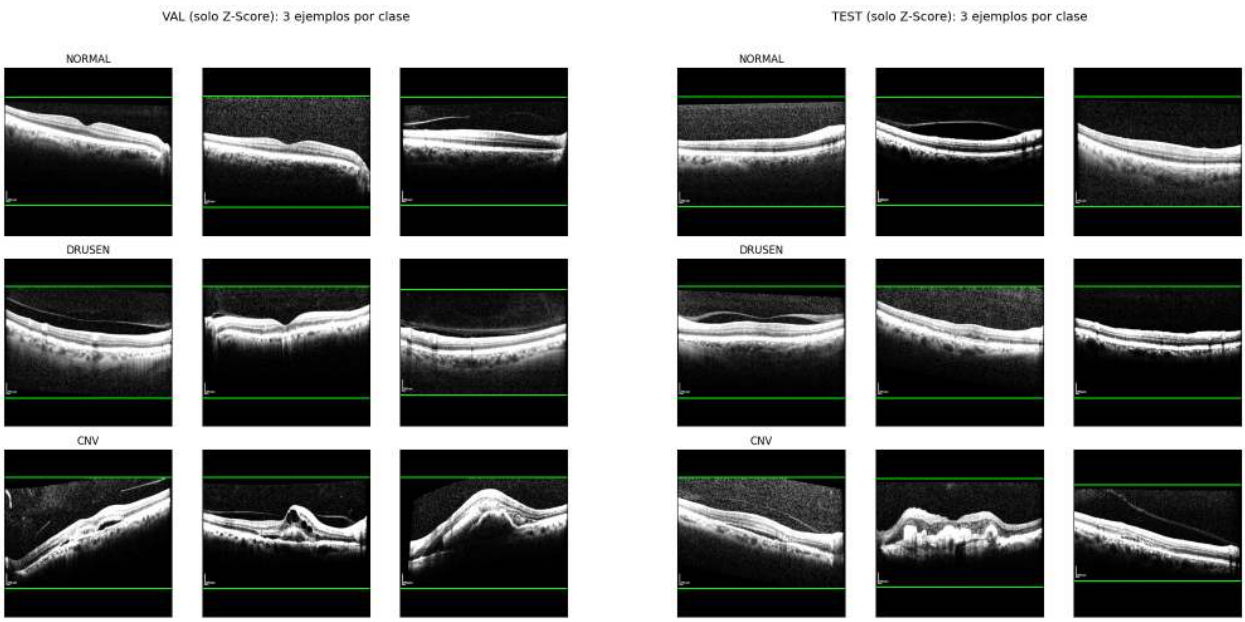


Figura 5.4: Visualización de muestras de validación y prueba externa con normalización *z-score*.

5.6.2. Variante 1: CSWin-Transformer + FPN.

La primera variante incorporó una **Feature Pyramid Network (FPN)** sobre las salidas jerárquicas del backbone. La FPN agrega mapas de características de diferentes etapas mediante conexiones laterales y un camino *top-down* con operaciones de *upsampling* y suma.

Cada rama lateral utiliza una proyección 1×1 para homogeneizar canales, seguida de una convolución 3×3 de suavizado tras la fusión. Este diseño permite combinar información semántica profunda y detalle espacial a múltiples escalas de representación.

5.6.3. Variante 2: CSWin-Transformer + ASPP.

La segunda variante sustituyó el módulo FPN por un bloque de **Atrous Spatial Pyramid Pooling (ASPP)**. Este módulo se acopló sobre un nivel intermedio del backbone y aplicó múltiples convoluciones paralelas con diferentes factores de dilatación para capturar contexto multiescala sin perder resolución espacial.

La configuración del ASPP incluyó:

- una convolución 1×1,
- tres convoluciones 3×3 con dilataciones $r \in \{6, 12, 18\}$,
- concatenación de salidas y proyección final 1×1.

A diferencia de FPN, que fusiona jerárquicamente representaciones de distintos niveles, ASPP explora múltiples campos receptivos efectivos dentro de una representación semánticamente rica.

5.6.4. Cabeza de clasificación.

En ambas variantes, los mapas producidos por el módulo multiescala fueron agregados mediante *pooling global* y proyectados posteriormente a una capa totalmente conectada de tres salidas correspondientes a *NORMAL*, *DRUSEN* y *CNV*.

De este modo, ambas arquitecturas compartieron:

- el mismo backbone CSWin-Transformer,
- el mismo preprocesamiento,
- la misma cabeza de clasificación,
- el mismo protocolo de evaluación,
- y un esquema general de optimización comparable.

Así, las diferencias observadas en desempeño pueden atribuirse fundamentalmente al módulo de agregación multiescala empleado.

5.7. Entrenamiento y validación de los modelos base.

5.7.1. Esquema general.

Las variantes CSWin-Transformer + FPN y CSWin-Transformer + ASPP se entrenaron bajo el mismo protocolo experimental de validación cruzada de 5 folds sobre el conjunto de desarrollo. Para cada fold se entrenó un modelo independiente y se seleccionó el mejor *checkpoint* correspondiente a partir de su desempeño en validación.

Posteriormente, cada modelo seleccionado fue evaluado sobre el mismo conjunto de prueba externo fijo. Las métricas finales reportadas en el capítulo de resultados corresponden al promedio y desviación estándar obtenidos al agregar los resultados de los cinco modelos evaluados sobre ese mismo conjunto de prueba externo.

5.7.2. Configuración de entrenamiento.

Los modelos se entrenaron para tres clases (*NORMAL*, *DRUSEN*, *CNV*) utilizando precisión mixta (*AMP*) en GPU, con tamaño de lote de 16 imágenes, `num_workers`=2 y `pin_memory` activado. El entrenamiento base se realizó durante **6 épocas** por fold, manteniendo la misma configuración general para ambas variantes dentro de cada serie comparativa.

5.7.3. Optimización y regularización.

Se utilizó el optimizador AdamW con decaimiento de peso:

$$\lambda = 5 \times 10^{-2}.$$

Se emplearon tasas de aprendizaje diferenciadas por bloque para preservar el conocimiento preentrenado del backbone y permitir una adaptación más rápida de los módulos superiores:

$$LR_{\text{backbone}} = 1 \times 10^{-4}, \qquad LR_{\text{neck}} = LR_{\text{head}} = 3 \times 10^{-4}.$$

La función de pérdida fue entropía cruzada multiclase. Para estabilizar el entrenamiento se aplicó *gradient clipping* con norma máxima de 1.0, así como precisión mixta con *autocast* y *GradScaler*. La tasa de aprendizaje se actualizó mediante un planificador *Cosine Annealing*.

Adicionalmente, durante el entrenamiento base se aplicó un **congelamiento parcial inicial del backbone** durante la primera época, manteniendo sin actualización los bloques tempranos *stage 1* y *stage 2*. Esta estrategia permitió estabilizar las representaciones de bajo nivel heredadas del preentrenamiento, mientras los módulos superiores de agregación multiescala y clasificación se adaptaban al dominio OCT. A partir de la segunda época, el entrenamiento continuó con el modelo habilitado de acuerdo con la configuración completa del experimento.

5.7.4. Bucle de entrenamiento y validación.

En cada fold, el entrenamiento se llevó a cabo recorriendo los subconjuntos de entrenamiento y validación correspondientes. En cada época se calcularon la pérdida media y la exactitud de ambos subconjuntos. Durante el entrenamiento se habilitó la retropropagación con precisión mixta, actualización de gradientes y paso del optimizador; durante validación se deshabilitaron gradientes para estimar el desempeño sin introducir sesgos.

Al finalizar cada época, se actualizó la tasa de aprendizaje mediante el planificador correspondiente.

5.7.5. Selección de checkpoints.

Para cada fold se almacenaron *checkpoints* de entrenamiento, incluyendo estado del modelo, optimizador, planificador, *GradScaler*, historial de métricas, época actual y mejor valor de validación. En el entrenamiento base, la selección del mejor *checkpoint* se realizó con base en la **exactitud de validación**, conservando además el último estado del entrenamiento y puntos de control intermedios para reanudación experimental.

5.8. Ajuste fino (*Fine-Tuning*) de las variantes.

Tras el entrenamiento base, se aplicó una segunda etapa de ajuste fino a ambas variantes arquitectónicas. El protocolo fue esencialmente el mismo para FPN y ASPP, adaptándose únicamente al módulo superior correspondiente en cada caso.

5.8.1. Esquema general de Fine-Tuning.

El ajuste fino partió del mejor *checkpoint* base obtenido en cada fold. En todos los casos se realizó un **fine-tuning corto de 3 épocas por fold**, con una estrategia de dos fases:

- **Época 1**: se mantuvo congelado todo el backbone CSWin-Transformer y se entrenaron exclusivamente el módulo superior y la cabeza clasificadora.
- **Épocas 2 y 3**: se descongeló únicamente el *stage 4* del backbone, manteniéndose congeladas las etapas anteriores, mientras continuó el ajuste del módulo superior y la cabeza clasificadora.

En la variante con FPN, el módulo superior ajustado correspondió a **FPN + cabeza clasificadora**; en la variante con ASPP, el ajuste se aplicó a **ASPP + cabeza clasificadora**.

Este diseño permitió preservar las representaciones generales aprendidas por el backbone y concentrar el refinamiento en las capas de mayor nivel semántico.

5.8.2. Optimización del Fine-Tuning.

El ajuste fino se realizó con AdamW y tasas de aprendizaje diferenciadas:

$$LR_{\text{backbone(stage 4)}} = 5 \times 10^{-6},$$

$$LR_{\text{neck}} = 3 \times 10^{-5},$$

$$LR_{\text{head}} = 5 \times 10^{-5}.$$

Se utilizó `CrossEntropyLoss` con pesos por clase calculados dinámicamente en cada fold y *label smoothing* de 0.02. Además, se empleó *gradient clipping* con norma máxima de 1.0 y un planificador *Cosine Annealing* a lo largo de las tres épocas.

5.8.3. Selección del mejor modelo ajustado.

En la etapa de fine-tuning, la selección del mejor *checkpoint* se realizó con base en el *macro-F1* de validación. Esta decisión se adoptó porque el problema es multiclase y la métrica macro-F1 resulta más sensible al equilibrio entre precisión y recall en las tres categorías.

5.9. Métricas de evaluación.

Dado que el problema es multiclase, presenta ligero desbalance y posee relevancia clínica en términos de falsos negativos, la evaluación del modelo no se limitó a la exactitud global. Se utilizaron las siguientes métricas:

5.9.1. Matriz de confusión.

La matriz de confusión resume la relación entre clase verdadera y clase predicha, permitiendo identificar patrones de error entre categorías diagnósticas. Para cada clase c se consideran:

$$TP_c, \quad FP_c, \quad TN_c, \quad FN_c.$$

5.9.2. Exactitud global.

La exactitud global se definió como:

$$Acc = \frac{\sum_c TP_c}{N},$$

donde N es el número total de muestras. Aunque resume el desempeño global, puede ocultar degradaciones específicas en clases con mayor complejidad diagnóstica.

5.9.3. Precisión, recall y F_1 por clase.

Para cada clase c se calcularon:

$$Prec_c = \frac{TP_c}{TP_c + FP_c}, \qquad Rec_c = \frac{TP_c}{TP_c + FN_c}, \qquad F_{1,c} = \frac{2\, Prec_c\, Rec_c}{Prec_c + Rec_c}.$$

La precisión penaliza falsos positivos, el recall penaliza falsos negativos y el F_1 resume el equilibrio entre ambas.

5.9.4. Promedios macro y ponderado.

Se reportaron métricas promedio macro y ponderado. El promedio macro asigna el mismo peso a cada clase y resulta útil para evaluar equilibrio interclase, mientras que el ponderado considera el soporte real de cada categoría.

5.9.5. Curvas ROC y AUC multiclase.

Para cada clase se construyeron curvas ROC bajo el esquema one-vs-rest (OvR), utilizando las probabilidades de salida del modelo. A partir de ellas se calcularon AUC por clase y AUC promedio macro:

$$AUC^{macro} = \frac{1}{C} \sum_c AUC_c.$$

El AUC permite evaluar la capacidad discriminativa del modelo independientemente del umbral de decisión y complementa la interpretación de las métricas discretas basadas en arg máx.

5.9.6. Soporte.

Se reportó también el *support*, entendido como el número de muestras verdaderas por clase. Esta métrica contextualiza la estabilidad e interpretación de precisión, recall y F_1, particularmente en clases con menor representación relativa.

5.10. Reproducibilidad experimental.

Con fines de trazabilidad y reproducibilidad, se conservaron los índices de la partición externa, los CSV de folds, los *checkpoints* por fold, la configuración de preprocesamiento, el mapeo de clases, las métricas de entrenamiento y validación, así como los resultados de evaluación sobre el conjunto de prueba externo. Esta organización permitió reconstruir tanto los modelos base como sus versiones ajustadas mediante fine-tuning bajo el mismo protocolo experimental.

CAPÍTULO 6

Resultados y discusión.

En este capitulo se analizan los resultados obtenidos al aplicar la metodología descrita anteriormente.

6.1. Protocolo de evaluación.

El protocolo común utilizado para la evaluación de todas las variantes fue el siguiente:

- **Datos**: *Labeled Retinal Optical Coherence Tomography Dataset for Classification of Normal, Drusen, and CNV Cases*, considerando las tres clases diagnósticas (NORMAL, DRUSEN y CNV) y manteniendo los mismos *splits* definidos para entrenamiento, validación y prueba externa.
- **Preprocesamiento**: padding simétrico, recorte de la región de interés (ROI), normalización z-score por imagen y augmentación aplicada únicamente en entrenamiento.
- **Backbone**: CSWin-Transformer adaptado a imágenes en escala de grises de un canal, con tamaño de entrada de 512×512.
- **Selección de modelo**: en cada variante se seleccionó el mejor *checkpoint* con base en el desempeño en validación, de acuerdo con la estrategia de entrenamiento correspondiente.
- **Evaluación en prueba**: cada modelo seleccionado fue evaluado sobre el mismo conjunto de prueba externo fijo.
- **Fine-Tuning**: cuando aplicó, se realizó como una segunda etapa de ajuste fino sobre los mejores modelos base obtenidos en cada fold.

Para problemas multiclase, el área bajo la curva ROC se estimó mediante el esquema one-vs-rest (OvR). Este enfoque permite evaluar la capacidad discriminativa de cada clase frente al resto, generando curvas ROC independientes para NORMAL, DRUSEN y CNV. Posteriormente, se reporta el promedio macro (Macro-AUC), el cual resume la separabilidad probabilística global del modelo sin verse afectado por el desbalance entre clases.

6.2. CSWin-Transformer + FPN.

6.2.1. Evaluación del modelo CSWin-Transformer + FPN.

En esta sección se presentan los resultados obtenidos con la arquitectura CSWin-Transformer incorporando una Feature Pyramid Network (FPN), evaluada bajo un protocolo de validación cruzada de 5 folds. Para cada fold se entrenó un modelo independiente y posteriormente se evaluó sobre el conjunto de prueba externo fijo, por lo que las métricas reportadas corresponden al promedio y desviación estándar obtenidos a partir de los cinco modelos.

Comportamiento del entrenamiento.

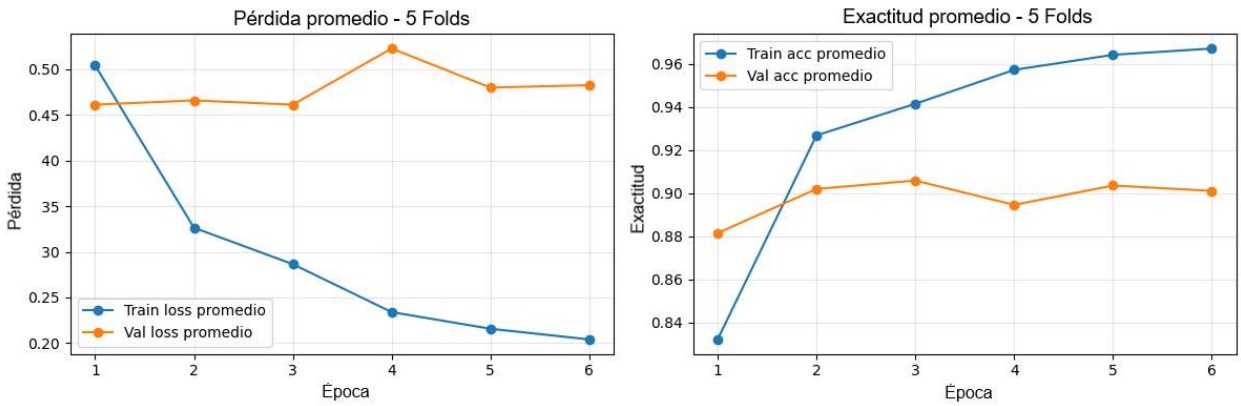


Figura 6.1: Gráficas de pérdida y exactitud promedio CSWin-Transformer+FPN.

La Figura 6.1 muestra el comportamiento promedio del entrenamiento y la validación a través de los cinco folds. En la subfigura a) se observa que la exactitud de entrenamiento incrementa de forma sostenida durante las primeras épocas, mientras que la exactitud de validación también presenta una mejora temprana y después tiende a estabilizarse. Este patrón indica que la arquitectura converge de manera consistente y que el proceso de aprendizaje es estable entre folds.

Por su parte, la subfigura b) muestra la evolución de la función de pérdida promedio. La pérdida de entrenamiento disminuye progresivamente, mientras que la pérdida de validación presenta una reducción inicial marcada y posteriormente fluctúa de forma más moderada. En conjunto, ambas curvas sugieren que el modelo logra aprender representaciones útiles para la clasificación sin evidenciar inestabilidad en el proceso de optimización.

Aunque hacia las últimas épocas se aprecia una separación moderada entre las curvas de entrenamiento y validación, dicha diferencia no sugiere un sobreajuste severo, sino un comportamiento esperado en una arquitectura de alta capacidad aplicada a imágenes OCT con patrones morfológicos complejos y, en ciertos casos, visualmente cercanos entre clases.

Métricas globales promedio.

Con el fin de resumir cuantitativamente el comportamiento general del modelo, en la Tabla 6.1 se presentan las métricas promedio obtenidas en los cinco folds sobre el conjunto de prueba externo.

Tabla 6.1: Métricas globales promedio del modelo CSWin-Transformer + FPN en el conjunto de prueba externo.

Métrica	**Valor**
Accuracy	0.9352 ± 0.0049
Precision macro	0.9394 ± 0.0039
Recall macro	0.9306 ± 0.0054
F1 macro	0.9336 ± 0.0045
Precision weighted	0.9366 ± 0.0042
Recall weighted	0.9352 ± 0.0049
F1 weighted	0.9344 ± 0.0051
Macro-AUC	0.9740 ± 0.0049
Micro-AUC	0.9766 ± 0.0052

La exactitud promedio alcanzada fue de 0.9352 ± 0.0049, lo que indica que el modelo clasifica correctamente, en promedio, el $93.52\,\%$ de las imágenes del conjunto de prueba externo. La baja desviación estándar asociada a esta métrica sugiere además un comportamiento estable entre folds.

El *precision macro* de 0.9394 ± 0.0039 y el *recall macro* de 0.9306 ± 0.0054 muestran que, al considerar las tres clases con el mismo peso, el modelo mantiene un desempeño alto tanto en la exactitud positiva de sus predicciones como en su capacidad de recuperación de casos verdaderos. La ligera disminución del recall respecto a la precisión sugiere que el modelo tiende a ser ligeramente más conservador en algunas clases, particularmente en aquellas con mayor dificultad diagnóstica.

El *F1 macro* de 0.9336 ± 0.0045 resulta especialmente relevante, ya que resume el equilibrio entre precisión y recall de manera no ponderada entre clases. Dado que esta métrica no favorece a la clase mayoritaria, su valor confirma que el desempeño del modelo es sólido incluso bajo una evaluación más estricta del balance interclase.

De manera complementaria, el *F1 weighted* de 0.9344 ± 0.0051 es muy cercano al F1 macro, lo que indica que el desbalance de clases no distorsiona de forma importante la interpretación global del rendimiento. Esta cercanía entre ambas métricas sugiere que el comportamiento del modelo es relativamente homogéneo y no depende exclusivamente del buen desempeño en una sola categoría.

Costo computacional y tiempos de ejecución.

Además del desempeño predictivo, resulta relevante analizar el costo computacional asociado al entrenamiento y evaluación del modelo. La Tabla 6.2 resume los tiempos promedio registrados durante los cinco folds.

Tabla 6.2: Tiempos promedio de entrenamiento e inferencia del modelo CSWin-Transformer + FPN en GPU A100.

Métrica temporal	**Valor**
Tiempo total de entrenamiento por fold	1179.0331 ± 14.9562 s
Tiempo total de validación por fold	88.1436 ± 6.2696 s
Tiempo total por fold	1268.0447 ± 9.7730 s
Tiempo de inferencia en test	18.7601 ± 0.6371 s
Tiempo de inferencia por imagen	7.4415 ± 0.2527 ms
Imágenes procesadas por segundo	134.4995 ± 4.3718
Tiempo promedio de entrenamiento por imagen	24.2551 ± 0.2107 ms
Tiempo promedio de validación por imagen	7.2527 ± 0.0147 ms

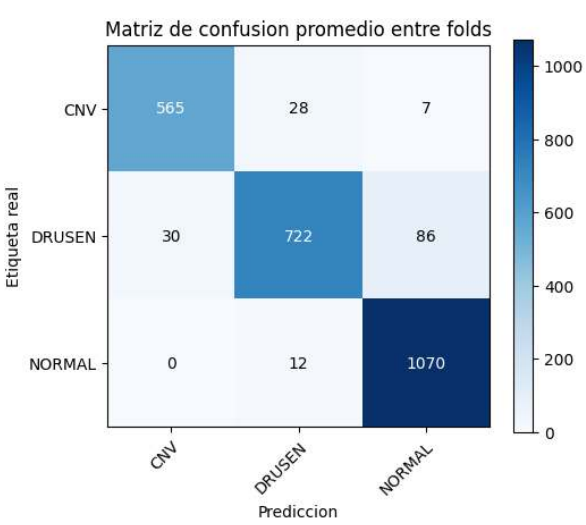
En promedio, cada fold requirió 1179.0331 ± 14.9562 segundos de entrenamiento y 88.1436 ± 6.2696 segundos de validación, lo que corresponde a un tiempo total de 1268.0447 ± 9.7730 segundos por fold. Expresado en unidades más interpretables, esto equivale aproximadamente a 19.65 minutos de entrenamiento puro y 21.13 minutos de ejecución total por fold. Considerando los cinco folds del protocolo experimental, el costo acumulado del proceso de entrenamiento resulta consistente con una estrategia de validación rigurosa y computacionalmente demandante.

En la etapa de inferencia sobre el conjunto de prueba externo, el modelo registró un tiempo promedio de 18.7601 ± 0.6371 segundos por evaluación completa, con un costo de 7.4415 ± 0.2527 ms por imagen y una velocidad aproximada de 134.4995 ± 4.3718 imágenes por segundo. Estos valores indican que, una vez entrenado, el modelo presenta una inferencia relativamente eficiente, incluso tratándose de una arquitectura basada en transformadores y procesamiento multiescala.

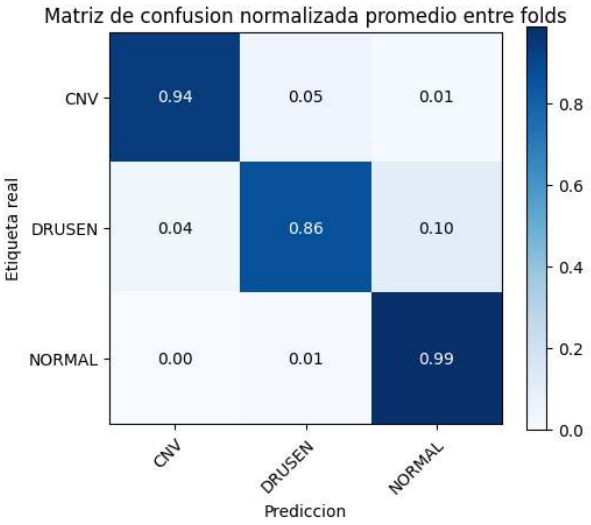
Estos tiempos deben interpretarse en el contexto del hardware utilizado, específicamente una GPU NVIDIA A100. La A100 proporciona una capacidad elevada de cómputo paralelo, gran ancho de banda de memoria y soporte eficiente para operaciones tensoriales aceleradas, lo cual favorece tanto el entrenamiento como la inferencia de modelos profundos de alta complejidad. En este sentido, los tiempos observados no sólo reflejan el costo computacional inherente de la arquitectura CSWin-Transformer + FPN, sino también el beneficio de haber ejecutado el experimento en una infraestructura de alto rendimiento.

Desde el punto de vista metodológico, esto significa que la GPU A100 permitió mantener tiempos de entrenamiento razonables a pesar de trabajar con validación cruzada de 5 folds, imágenes OCT de alta resolución y una arquitectura con backbone transformer. Asimismo, la velocidad de inferencia alcanzada confirma que el modelo puede procesar el conjunto de prueba completo en pocos segundos, lo que refuerza su viabilidad operativa en escenarios de análisis automatizado. No obstante, debe considerarse que estos tiempos dependen directamente de la configuración experimental empleada, incluyendo tamaño de lote, número de épocas, estrategia de carga de datos y recursos de hardware disponibles.

Análisis de la matriz de confusión.



(a) Matriz de confusión CSWin+FPN.



(b) Matriz de confusión normalizada CSWin+FPN.

La matriz de confusión promedio mostrada en la Figura 6.2a resume el comportamiento del modelo sobre el conjunto de prueba externo al promediar los resultados de los cinco folds. En términos de conteos promedio, la matriz indica aproximadamente 565 casos de CNV correctamente clasificados, 722.4 casos de DRUSEN correctamente identificados y 1070.2 imágenes NORMAL reconocidas de manera adecuada.

La matriz normalizada de la Figura 6.2b permite analizar con mayor claridad la sensibilidad promedio de cada clase. Los recalls promedio obtenidos fueron de 0.9417 ± 0.0213 para CNV, 0.8610 ± 0.0241 para DRUSEN y 0.9891 ± 0.0022 para NORMAL. Estos resultados muestran que la clase NORMAL es la mejor recuperada por el modelo, mientras que DRUSEN representa la categoría de mayor dificultad relativa.

El patrón de error observado indica que la principal fuente de confusión se concentra en la clase DRUSEN. En promedio, esta clase presenta confusiones tanto hacia CNV como hacia NORMAL, aunque la transición hacia CNV es la más importante. Este comportamiento es clínicamente razonable, ya que ciertas manifestaciones estructurales de DRUSEN pueden compartir rasgos morfológicos con otras alteraciones maculares, dificultando su separación automática.

Métricas promedio por clase.

Para profundizar en el análisis interclase, la Tabla 6.3 presenta las métricas promedio por clase obtenidas en el conjunto de prueba externo.

Tabla 6.3: Métricas promedio por clase del modelo CSWin-Transformer + FPN en el conjunto de prueba externo.

Clase	Precisión	Recall	F1-score	AUC
CNV	0.9499 ± 0.0248	0.9417 ± 0.0213	0.9453 ± 0.0025	0.9745 ± 0.0082
DRUSEN	0.9485 ± 0.0139	0.8610 ± 0.0241	0.9023 ± 0.0083	0.9624 ± 0.0102
NORMAL	0.9199 ± 0.0140	0.9891 ± 0.0022	0.9532 ± 0.0068	0.9842 ± 0.0054

La clase **CNV** presenta un desempeño alto y equilibrado, con precisión de 0.9499 ± 0.0248, recall de 0.9417 ± 0.0213 y F1-score de 0.9453 ± 0.0025. Esto indica que el modelo no sólo identifica correctamente una gran proporción de los casos reales de CNV, sino que además mantiene una baja tasa de falsos positivos al asignar esta clase.

En **DRUSEN**, la precisión se mantiene elevada (0.9485 ± 0.0139), lo que sugiere que cuando el modelo predice esta categoría suele hacerlo correctamente. Sin embargo, el recall disminuye a 0.8610 ± 0.0241, evidenciando que una fracción importante de casos reales de DRUSEN no es recuperada y se confunde con otras clases. Esta diferencia entre precisión y recall explica que el F1-score de DRUSEN (0.9023 ± 0.0083) sea el más bajo de las tres categorías, confirmando que esta clase constituye el principal desafío para la arquitectura.

La clase **NORMAL** alcanza el recall más alto del estudio (0.9891 ± 0.0022), lo que indica que prácticamente todos los casos normales son identificados correctamente. Aunque su precisión (0.9199 ± 0.0140) es menor que la observada en CNV y DRUSEN, su F1-score permanece elevado (0.9532 ± 0.0068). Esto sugiere que el modelo muestra una excelente capacidad para detectar normalidad, aunque en algunos casos asigna esta categoría a imágenes patológicas que comparten rasgos menos evidentes.

Curvas ROC y capacidad discriminativa.

Las curvas ROC promedio bajo el esquema One-vs-Rest, mostradas en la Figura 6.3, permiten analizar la capacidad discriminativa del modelo de forma independiente para cada clase. Los valores promedio de AUC fueron de 0.9745 ± 0.0082 para CNV, 0.9624 ± 0.0102 para DRUSEN y 0.9842 ± 0.0054 para NORMAL.

Estos resultados indican que el modelo posee una alta capacidad de separación entre clases incluso cuando se consideran distintos umbrales de decisión. En particular, la clase NORMAL presenta la mayor separabilidad global, mientras que DRUSEN vuelve a mostrar el valor más bajo, en concordancia con lo observado previamente en la matriz de confusión y en las métricas de recuperación.

A nivel global, el *macro-AUC* alcanzó 0.9740 ± 0.0049 y el *micro-AUC* fue de 0.9766 ± 0.0052. El macro-AUC resulta especialmente importante porque refleja la capacidad discriminativa promedio tratando a todas las clases por igual, mientras que el micro-AUC considera el conjunto total de predicciones. La cercanía entre ambos valores indica que el modelo conserva una separabilidad alta y relativamente uniforme en todo el problema de clasificación multiclase.

Síntesis del desempeño.

En conjunto, el modelo CSWin-Transformer + FPN demuestra:

- Convergencia estable durante el entrenamiento promedio de los 5 folds.

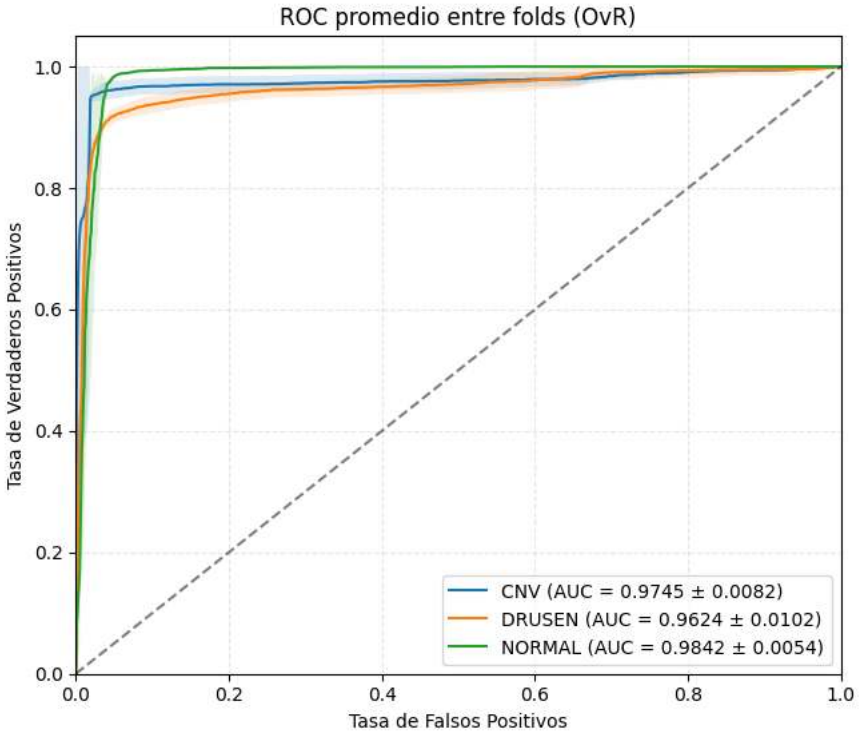


Figura 6.3: Curvas ROC de las 3 clases.

- Exactitud global de 0.9352 ± 0.0049.
- Precision macro de 0.9394 ± 0.0039, recall macro de 0.9306 ± 0.0054 y F1 macro de 0.9336 ± 0.0045.
- Weighted-F1 de 0.9344 ± 0.0051, consistente con el desempeño global del modelo.
- Macro-AUC de 0.9740 ± 0.0049 y Micro-AUC de 0.9766 ± 0.0052.
- Tiempo promedio de entrenamiento de 1179.0331 ± 14.9562 s por fold e inferencia de 18.7601 ± 0.6371 s sobre el conjunto de prueba externo.
- Velocidad de inferencia de 134.4995 ± 4.3718 imágenes por segundo en GPU A100.
- Desempeño alto en las tres clases, con mayor dificultad relativa en la identificación de DRUSEN.

6.3. CSWin-Transformer + ASPP

6.3.1. Evaluación del modelo CSWin-Transformer + ASPP.

En esta sección se presentan los resultados obtenidos con la arquitectura CSWin-Transformer incorporando el módulo Atrous Spatial Pyramid Pooling (ASPP), evaluada bajo un protocolo de

validación cruzada de 5 folds. Para cada fold se entrenó un modelo independiente y posteriormente se evaluó sobre el conjunto de prueba externo fijo, por lo que las métricas reportadas corresponden al promedio y desviación estándar obtenidos a partir de los cinco modelos.

Comportamiento del entrenamiento

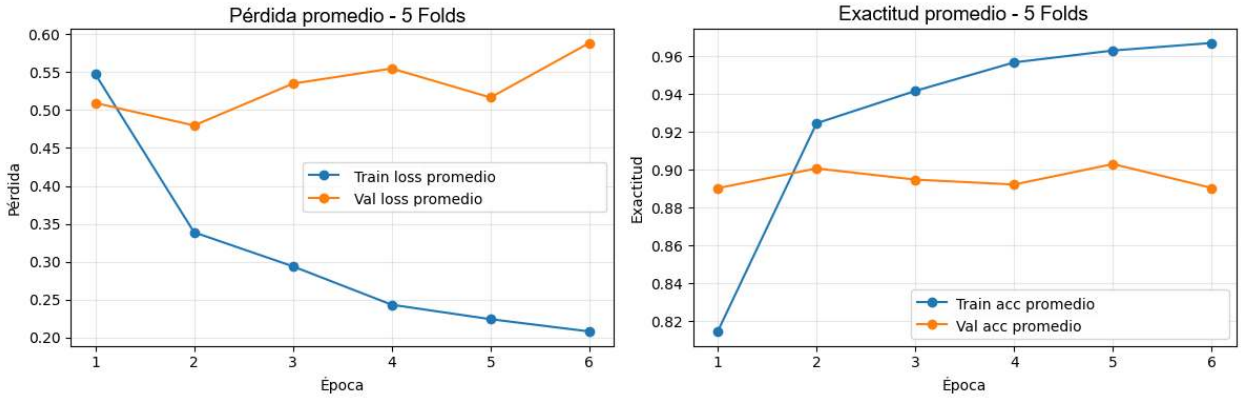


Figura 6.4: Gráficas de pérdida y exactitud promedio CSWin-Transformer+ASPP.

La Figura 6.4 muestra el comportamiento promedio del entrenamiento y la validación a través de los cinco folds. En la subfigura a) se observa que la exactitud de entrenamiento incrementa de forma sostenida durante las primeras épocas, mientras que la exactitud de validación también presenta una mejora temprana y posteriormente tiende a estabilizarse. Este comportamiento indica que la arquitectura converge de forma consistente y que el proceso de aprendizaje se mantiene estable entre folds.

Por su parte, la subfigura b) presenta la evolución promedio de la función de pérdida. La pérdida de entrenamiento disminuye progresivamente, mientras que la pérdida de validación muestra una reducción importante en las primeras épocas y posteriormente presenta oscilaciones moderadas sin incrementos abruptos. En conjunto, ambas curvas reflejan un entrenamiento controlado y una optimización estable del modelo.

Aunque hacia las últimas épocas se aprecia una separación moderada entre entrenamiento y validación, esta diferencia no sugiere un sobreajuste severo, sino un comportamiento esperable en una arquitectura de alta capacidad aplicada a imágenes OCT con patrones morfológicos complejos y similitud visual entre ciertas clases.

Métricas globales promedio.

Con el fin de resumir cuantitativamente el comportamiento general del modelo, en la Tabla 6.4 se presentan las métricas promedio obtenidas en los cinco folds sobre el conjunto de prueba externo.

Tabla 6.4: Métricas globales promedio del modelo CSWin-Transformer + ASPP en el conjunto de prueba externo.

Métrica	**Valor**
Accuracy	0.9440 ± 0.0073
Precision macro	0.9470 ± 0.0049
Recall macro	0.9408 ± 0.0090
F1 macro	0.9433 ± 0.0069
Precision weighted	0.9445 ± 0.0068
Recall weighted	0.9440 ± 0.0073
F1 weighted	0.9436 ± 0.0072
Macro-AUC	0.9799 ± 0.0014
Micro-AUC	0.9811 ± 0.0011

La exactitud promedio alcanzada fue de 0.9440 ± 0.0073, lo que indica que el modelo clasifica correctamente, en promedio, el 94.40 % de las imágenes del conjunto de prueba externo. La desviación estándar relativamente baja muestra que el comportamiento del modelo se mantiene estable entre los cinco folds evaluados.

El *precision macro* de 0.9470 ± 0.0049 y el *recall macro* de 0.9408 ± 0.0090 evidencian que, al ponderar por igual las tres clases, la arquitectura conserva un desempeño alto tanto en la exactitud positiva de sus predicciones como en la recuperación de casos verdaderos. La diferencia moderada entre ambas métricas sugiere que, aunque el modelo presenta buena capacidad global de clasificación, ciertas clases siguen siendo más exigentes desde el punto de vista de la sensibilidad.

El *F1 macro* de 0.9433 ± 0.0069 resulta particularmente relevante porque resume el equilibrio entre precisión y recall sin favorecer a la clase mayoritaria. Su valor confirma que el modelo mantiene un rendimiento sólido en el problema multiclase, incluso bajo un criterio de evaluación estricto respecto al balance interclase.

Por otra parte, el *F1 weighted* de 0.9436 ± 0.0072 es muy cercano al F1 macro, lo que indica que el desbalance del conjunto no altera de manera importante la interpretación del desempeño global. Esta proximidad entre ambas métricas sugiere que la arquitectura ASPP ofrece una respuesta relativamente homogénea entre clases.

Costo computacional y tiempos de ejecución.

Además del desempeño predictivo, resulta importante analizar el costo computacional asociado al entrenamiento y evaluación del modelo. La Tabla 6.5 resume los tiempos promedio registrados durante los cinco folds.

Tabla 6.5: Tiempos promedio de entrenamiento e inferencia del modelo CSWin-Transformer + ASPP en GPU A100.

Métrica temporal	**Valor**
Tiempo total de entrenamiento por fold	1169.9877 ± 16.9603 s
Tiempo total de validación por fold	83.1835 ± 5.9078 s
Tiempo total por fold	1256.2208 ± 15.1983 s
Tiempo de inferencia en test	17.4758 ± 0.0997 s
Tiempo de inferencia por imagen	6.9321 ± 0.0395 ms
Imágenes procesadas por segundo	144.2607 ± 0.8209
Tiempo promedio de entrenamiento por imagen	24.0683 ± 0.1791 ms
Tiempo promedio de validación por imagen	6.8446 ± 0.0139 ms

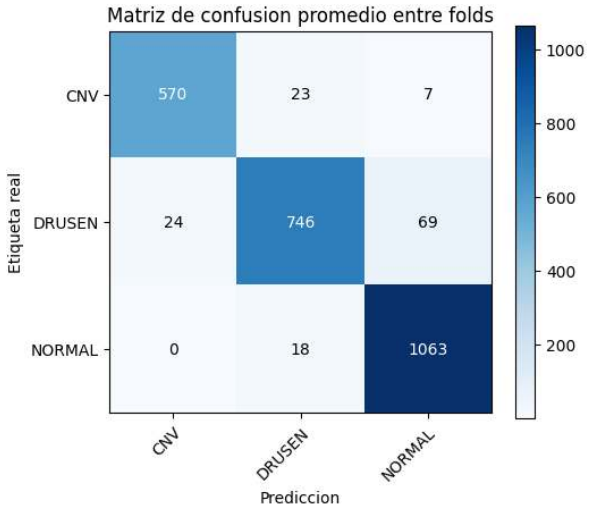
En promedio, cada fold requirió 1169.9877 ± 16.9603 segundos de entrenamiento y 83.1835 ± 5.9078 segundos de validación, lo que corresponde a un tiempo total de 1256.2208 ± 15.1983 segundos por fold. Expresado en unidades más interpretables, esto equivale aproximadamente a 19.50 minutos de entrenamiento puro y 20.94 minutos de ejecución total por fold. Considerando los cinco folds del protocolo experimental, el costo acumulado del entrenamiento es consistente con una estrategia metodológicamente rigurosa y computacionalmente exigente.

En la etapa de inferencia sobre el conjunto de prueba externo, el modelo registró un tiempo promedio de 17.4758 ± 0.0997 segundos por evaluación completa, con un costo de 6.9321 ± 0.0395 ms por imagen y una velocidad aproximada de 144.2607 ± 0.8209 imágenes por segundo. Estos valores indican que, una vez entrenado, el modelo presenta una inferencia eficiente incluso tratándose de una arquitectura basada en transformadores y un módulo multiescala como ASPP.

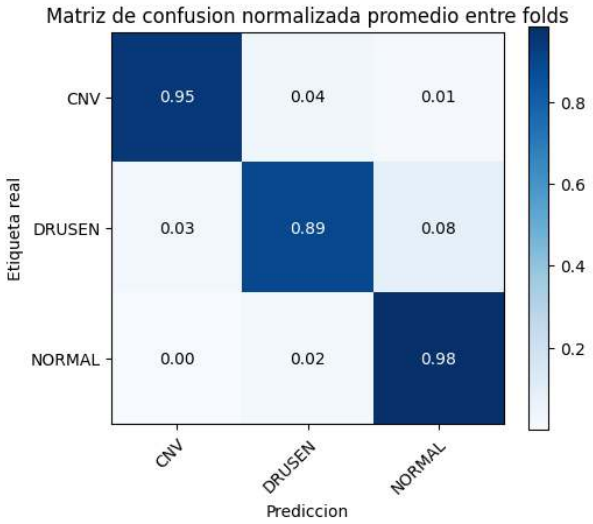
Estos tiempos deben interpretarse en el contexto del hardware empleado, específicamente una GPU NVIDIA A100. La A100 proporciona una elevada capacidad de cómputo paralelo, alto ancho de banda de memoria y soporte eficiente para operaciones tensoriales aceleradas, lo que favorece tanto el entrenamiento como la inferencia de modelos profundos de alta complejidad. En consecuencia, los tiempos observados reflejan no sólo el costo computacional de la arquitectura CSWin-Transformer + ASPP, sino también el beneficio de haber ejecutado los experimentos sobre una infraestructura de alto rendimiento.

Desde una perspectiva metodológica, esto implica que la GPU A100 permitió mantener tiempos de entrenamiento razonables aun trabajando con validación cruzada de 5 folds, imágenes OCT de alta resolución y una arquitectura con backbone transformer. Asimismo, la velocidad de inferencia obtenida confirma que el modelo puede procesar el conjunto de prueba completo en pocos segundos, lo que refuerza su factibilidad operativa en escenarios de apoyo al análisis automatizado. No obstante, debe considerarse que estos tiempos dependen directamente de la configuración experimental empleada, incluyendo tamaño de lote, número de épocas, estrategia de carga de datos y recursos de hardware disponibles.

Análisis de la matriz de confusión.



(a) Matriz de confusión CSWin+ASPP.



(b) Matriz de confusión normalizada CSWin+ASPP.

La matriz de confusión promedio mostrada en la Figura 6.5a resume el comportamiento del modelo sobre el conjunto de prueba externo al promediar los resultados de los cinco folds. En términos de conteos promedio, la matriz indica aproximadamente 570.2 casos de CNV correctamente clasificados, 745.8 casos de DRUSEN correctamente identificados y 1063.4 imágenes NORMAL reconocidas de manera adecuada.

La matriz normalizada de la Figura 6.5b permite analizar con mayor claridad la sensibilidad promedio de cada clase. Los recalls promedio obtenidos fueron de 0.9503 ± 0.0195 para CNV, 0.8894 ± 0.0080 para DRUSEN y 0.9828 ± 0.0042 para NORMAL. Estos resultados muestran que la clase NORMAL sigue siendo la mejor recuperada por el modelo, mientras que DRUSEN continúa representando la categoría de mayor dificultad relativa.

El patrón de error observado indica que la principal fuente de confusión se concentra nuevamente en la clase DRUSEN. Aunque la arquitectura ASPP mejora el equilibrio global del modelo, esta categoría sigue mostrando mayor susceptibilidad a confundirse con las restantes clases, lo que sugiere que sus rasgos discriminativos continúan siendo más complejos de modelar que en CNV o NORMAL.

Métricas promedio por clase.

Para profundizar en el análisis interclase, la Tabla 6.6 presenta las métricas promedio por clase obtenidas en el conjunto de prueba externo.

Tabla 6.6: Métricas promedio por clase del modelo CSWin-Transformer + ASPP en el conjunto de prueba externo.

Clase	Precisión	Recall	F1-score	AUC
CNV	0.9594 ± 0.0164	0.9503 ± 0.0195	0.9546 ± 0.0045	0.9850 ± 0.0033
DRUSEN	0.9478 ± 0.0142	0.8894 ± 0.0080	0.9176 ± 0.0104	0.9671 ± 0.0031
NORMAL	0.9338 ± 0.0140	0.9828 ± 0.0042	0.9576 ± 0.0071	0.9868 ± 0.0032

La clase **CNV** presenta un desempeño alto y equilibrado, con precisión de 0.9594 ± 0.0164, recall de 0.9503 ± 0.0195 y F1-score de 0.9546 ± 0.0045. Esto indica que el modelo identifica correctamente una gran proporción de los casos reales de CNV y, además, mantiene una baja tasa de falsos positivos al asignar esta categoría.

En **DRUSEN**, la precisión permanece alta (0.9478 ± 0.0142), lo que sugiere que cuando el modelo predice esta clase suele hacerlo correctamente. Sin embargo, el recall disminuye a 0.8894 ± 0.0080, lo que indica que una fracción de los casos reales de DRUSEN continúa siendo confundida con otras categorías. Esta diferencia entre precisión y recall explica que el F1-score de DRUSEN (0.9176 ± 0.0104) sea el más bajo entre las tres clases, confirmando que esta categoría sigue siendo la más desafiante para la arquitectura.

La clase **NORMAL** alcanza el recall más alto del modelo (0.9828 ± 0.0042), lo que indica que la gran mayoría de los casos normales son detectados correctamente. Aunque su precisión (0.9338 ± 0.0140) es menor que la observada en CNV, su F1-score se mantiene elevado (0.9576 ± 0.0071). Esto sugiere que la arquitectura presenta una muy buena capacidad para identificar normalidad, aunque en algunos casos todavía asigna esta categoría a imágenes patológicas con características menos marcadas.

Curvas ROC y capacidad discriminativa.

Las curvas ROC promedio bajo el esquema One-vs-Rest, mostradas en la Figura 6.6, permiten analizar la capacidad discriminativa del modelo de forma independiente para cada clase. Los valores promedio de AUC fueron de 0.9850 ± 0.0033 para CNV, 0.9671 ± 0.0031 para DRUSEN y 0.9868 ± 0.0032 para NORMAL.

Estos resultados indican que el modelo posee una alta capacidad de separación entre clases incluso al considerar distintos umbrales de decisión. En particular, la clase NORMAL presenta la mayor separabilidad global, seguida muy de cerca por CNV, mientras que DRUSEN vuelve a mostrar el valor más bajo, en concordancia con lo observado previamente en la matriz de confusión y en las métricas de recuperación.

A nivel global, el *macro-AUC* alcanzó 0.9799 ± 0.0014 y el *micro-AUC* fue de 0.9811 ± 0.0011. El macro-AUC resulta especialmente importante porque refleja la capacidad discriminativa promedio tratando a todas las clases por igual, mientras que el micro-AUC considera el conjunto total de predicciones. La cercanía entre ambos valores indica que el modelo conserva una separabilidad alta y relativamente uniforme en todo el problema de clasificación multiclase.

Síntesis del desempeño.

El modelo CSWin-Transformer + ASPP presenta:

- Convergencia estable durante el entrenamiento promedio de los 5 folds.

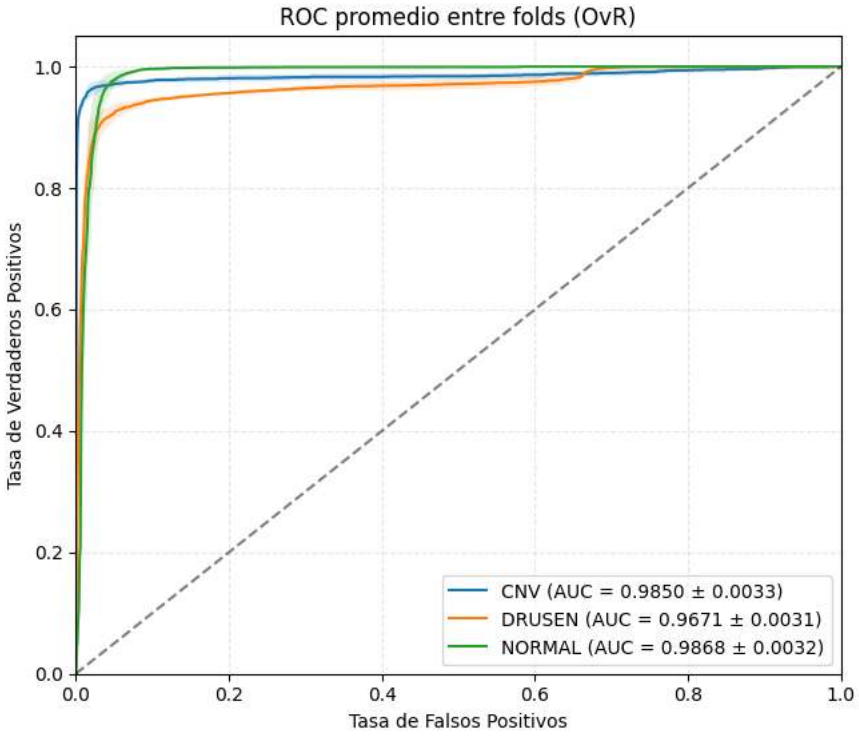


Figura 6.6: Curvas ROC de las 3 clases.

- Exactitud global de 0.9440 ± 0.0073.
- Precision macro de 0.9470 ± 0.0049, recall macro de 0.9408 ± 0.0090 y F1 macro de 0.9433 ± 0.0069.
- Weighted-F1 de 0.9436 ± 0.0072, consistente con el desempeño global del modelo.
- Macro-AUC de 0.9799 ± 0.0014 y Micro-AUC de 0.9811 ± 0.0011.
- Tiempo promedio de entrenamiento de 1169.9877 ± 16.9603 s por fold e inferencia de 17.4758 ± 0.0997 s sobre el conjunto de prueba externo.
- Velocidad de inferencia de 144.2607 ± 0.8209 imágenes por segundo en GPU A100.
- Desempeño alto en las tres clases, con mayor dificultad relativa en la identificación de DRUSEN.

6.4. Comparación: CSWin-Transformer + FPN vs CSWin-Transformer + ASPP.

Con el fin de comparar de forma integral ambas variantes propuestas, la Tabla 6.7 resume las métricas globales promedio obtenidas por los modelos CSWin-Transformer + FPN y CSWin-

Transformer + ASPP bajo el protocolo de validación cruzada de 5 folds y evaluación sobre el conjunto de prueba externo.

Tabla 6.7: Comparación de desempeño entre CSWin-Transformer + FPN y CSWin-Transformer + ASPP

Métrica	**CSWin + FPN**	**CSWin + ASPP**
Accuracy	0.9352 ± 0.0049	$\mathbf{0.9440 \pm 0.0073}$
Precision macro	0.9394 ± 0.0039	$\mathbf{0.9470 \pm 0.0049}$
Recall macro	0.9306 ± 0.0054	$\mathbf{0.9408 \pm 0.0090}$
Macro-F1	0.9336 ± 0.0045	$\mathbf{0.9433 \pm 0.0069}$
Weighted-F1	0.9344 ± 0.0051	$\mathbf{0.9436 \pm 0.0072}$
Macro-AUC (OvR)	0.9740 ± 0.0049	$\mathbf{0.9799 \pm 0.0014}$
Micro-AUC (OvR)	0.9766 ± 0.0052	$\mathbf{0.9811 \pm 0.0011}$

La Tabla 6.7 evidencia que la variante CSWin-Transformer + ASPP supera de manera consistente a CSWin-Transformer + FPN en todas las métricas globales evaluadas. En términos de exactitud, ASPP alcanza 0.9440 ± 0.0073, superando el valor de 0.9352 ± 0.0049 obtenido por FPN. Esta diferencia indica una mejora global en la proporción de imágenes correctamente clasificadas en el conjunto de prueba externo.

La misma tendencia se observa en las métricas de balance interclase. El *Macro-F1* aumenta de 0.9336 ± 0.0045 con FPN a 0.9433 ± 0.0069 con ASPP, mientras que el *Weighted-F1* pasa de 0.9344 ± 0.0051 a 0.9436 ± 0.0072. Dado que el F1-score integra simultáneamente precisión y recall, estos resultados sugieren que ASPP no sólo mejora el desempeño global, sino que también ofrece una respuesta más equilibrada entre las tres categorías diagnósticas.

La ventaja de ASPP también se refleja en la capacidad discriminativa del modelo. El *Macro-AUC* se incrementa de 0.9740 ± 0.0049 a 0.9799 ± 0.0014, mientras que el *Micro-AUC* mejora de 0.9766 ± 0.0052 a 0.9811 ± 0.0011. Estos incrementos indican una mayor separación entre distribuciones de probabilidad y, por tanto, una mejor capacidad del modelo para distinguir entre clases al considerar distintos umbrales de decisión. Además, la menor desviación estándar observada en ASPP para ambas métricas AUC sugiere un comportamiento más estable entre folds en términos de separabilidad.

Para complementar la comparación de desempeño, la Tabla 6.8 presenta los tiempos promedio de entrenamiento e inferencia registrados para ambas arquitecturas en GPU NVIDIA A100.

Tabla 6.8: Comparación de tiempos de ejecución entre CSWin-Transformer + FPN y CSWin-Transformer + ASPP en GPU A100

Métrica temporal	**CSWin + FPN**	**CSWin + ASPP**
Tiempo total de entrenamiento por fold	1179.0331 ± 14.9562 s	$\mathbf{1169.9877 \pm 16.9603}$ s
Tiempo total de validación por fold	88.1436 ± 6.2696 s	$\mathbf{83.1835 \pm 5.9078}$ s
Tiempo total por fold	1268.0447 ± 9.7730 s	$\mathbf{1256.2208 \pm 15.1983}$ s
Tiempo de inferencia en test	18.7601 ± 0.6371 s	$\mathbf{17.4758 \pm 0.0997}$ s
Tiempo de inferencia por imagen	7.4415 ± 0.2527 ms	$\mathbf{6.9321 \pm 0.0395}$ ms
Imágenes procesadas por segundo	134.4995 ± 4.3718	$\mathbf{144.2607 \pm 0.8209}$

Desde el punto de vista computacional, ASPP también presenta una ligera ventaja sobre FPN. El tiempo total de entrenamiento por fold es marginalmente menor en ASPP (1169.9877 ± 16.9603 s) que en FPN (1179.0331 ± 14.9562 s), y la inferencia sobre el conjunto de prueba externo también es más rápida, con 17.4758 ± 0.0997 s frente a 18.7601 ± 0.6371 s. En términos de rendimiento por imagen, ASPP reduce el tiempo de inferencia de 7.4415 ± 0.2527 ms a 6.9321 ± 0.0395 ms y aumenta la velocidad de procesamiento de 134.4995 ± 4.3718 a 144.2607 ± 0.8209 imágenes por segundo.

Estos resultados adquieren mayor relevancia al considerar que ambos modelos fueron entrenados y evaluados en una GPU NVIDIA A100. La elevada capacidad de cómputo paralelo de esta unidad permitió ejecutar un protocolo de validación cruzada de 5 folds con arquitecturas basadas en transformadores y módulos multiescala en tiempos razonables. Sin embargo, aun bajo las mismas condiciones de hardware, ASPP mostró una relación ligeramente más favorable entre desempeño predictivo y costo computacional, lo que fortalece su posición como la alternativa más eficiente dentro del presente estudio.

Si se analizan las diferencias arquitectónicas, ambas configuraciones parten del backbone CSWin-Transformer, el cual genera representaciones jerárquicas y multiescala mediante atención cruzada en ventanas. No obstante, el modo en que cada variante explota dichas representaciones es distinto.

La FPN integra la información de diferentes escalas por medio de una fusión piramidal jerárquica, combinando mapas de características de distinta resolución mediante conexiones laterales y un flujo *top-down*. Este mecanismo favorece la incorporación conjunta de información semántica y espacial distribuida a lo largo de varias escalas de representación.

En contraste, el módulo ASPP aplica convoluciones dilatadas en paralelo con distintas tasas de dilatación sobre una representación de alto nivel. Esta estrategia permite ampliar el campo receptivo efectivo sin perder resolución espacial, capturando contexto multiescala directamente sobre mapas con alto contenido semántico. En términos prácticos, ASPP facilita la integración simultánea de:

- patrones locales finos,
- alteraciones estructurales intermedias,
- y contexto morfológico más amplio dentro de la retina.

En imágenes OCT, esta diferencia resulta especialmente relevante, ya que clases como DRUSEN y CNV pueden presentar manifestaciones tanto focales como extendidas. Mientras FPN enfatiza la fusión jerárquica de escalas provenientes de diferentes niveles del backbone, ASPP amplía la percepción contextual sobre una representación semánticamente rica. Esta propiedad puede favorecer la detección de alteraciones cuya discriminación depende no sólo de un patrón local aislado, sino también de su relación con el entorno anatómico circundante.

Este comportamiento se refleja de manera particularmente clara en la clase DRUSEN. Aunque en ambas arquitecturas esta categoría continúa siendo la más difícil de clasificar, ASPP mejora su desempeño respecto a FPN. En términos de recall, DRUSEN pasa de 0.8610 ± 0.0241 con FPN a 0.8894 ± 0.0080 con ASPP, mientras que su F1-score aumenta de 0.9023 ± 0.0083 a 0.9176 ± 0.0104. De forma paralela, el AUC de esta clase se incrementa de 0.9624 ± 0.0102 a 0.9671 ± 0.0031. Estos resultados sugieren que ASPP logra modelar con mayor eficacia la variabilidad morfológica asociada a DRUSEN, reduciendo parte de la confusión observada con las otras clases.

En síntesis, ambas variantes confirman que el backbone CSWin-Transformer constituye una base sólida para la clasificación de imágenes OCT. No obstante, la incorporación del módulo ASPP ofrece ventajas consistentes frente a FPN al proporcionar:

- mejores métricas globales de clasificación,
- mayor capacidad discriminativa,
- mejor recuperación de la clase DRUSEN,
- y una inferencia ligeramente más eficiente en GPU A100.

Por ello, dentro del conjunto de experimentos realizados, CSWin-Transformer + ASPP se posiciona como la variante con mejor compromiso entre precisión diagnóstica, estabilidad entre folds y eficiencia computacional.

6.5. Ablación de Fine-Tuning: FPN vs ASPP.

Al haber realizado el Fine-Tuning de ambos modelos, en esta sección se discuten los resultados obtenidos.

6.5.1. Impacto en CSWin+FPN.

Estrategia en CSWin-Transformer + FPN.

El ajuste fino del modelo CSWin-Transformer + FPN se realizó como una segunda etapa de optimización incremental a partir de los mejores *checkpoints* obtenidos en el entrenamiento base para cada uno de los 5 folds. En esta fase no se introdujeron cambios en la arquitectura, por lo que se mantuvieron intactos el backbone CSWin-Transformer, el módulo FPN y la cabeza de clasificación asociada a la representación multiescala.

A diferencia del entrenamiento base, en esta etapa se aplicó una estrategia de fine-tuning corta y controlada, diseñada para refinar selectivamente las representaciones ya aprendidas sin reentrenar de forma extensa todo el modelo. Para ello, el ajuste fino se ejecutó durante 3 épocas por fold, utilizando como punto de partida el mejor modelo base correspondiente a cada partición.

La estrategia de actualización de parámetros se organizó en dos fases. Durante la primera época, únicamente se entrenaron los parámetros del módulo FPN y de la cabeza clasificadora, manteniendo congelado el backbone. A partir de la segunda época, se habilitó el entrenamiento del *stage 4* del backbone, mientras que los niveles anteriores permanecieron congelados. Este esquema permitió conservar las representaciones profundas ya estabilizadas en las etapas tempranas del backbone y concentrar el ajuste sobre las capas de más alto nivel semántico, que son las más relevantes para la adaptación final al dominio OCT.

La optimización se realizó con AdamW y tasas de aprendizaje diferenciadas por bloque:

- Backbone CSWin-Transformer: 5×10^{-6}
- FPN: 3×10^{-5}
- Cabeza clasificadora: 5×10^{-5}

Esta configuración favorece un ajuste conservador del backbone y una mayor plasticidad en los módulos superiores responsables de la integración multiescala y la decisión final de clasificación.

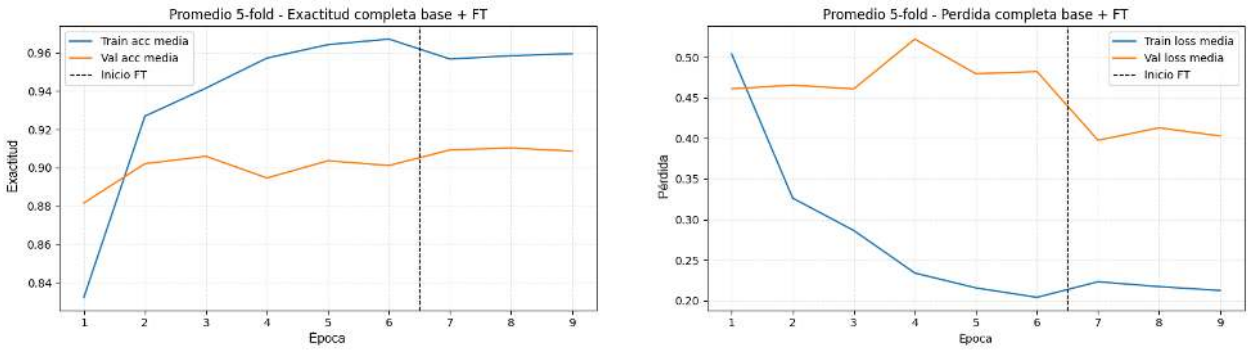
Como función de pérdida se empleó `CrossEntropyLoss` con pesos por clase calculados dinámicamente en cada fold y *label smoothing* de 0.02. La incorporación de pesos de clase permitió compensar diferencias de frecuencia entre categorías dentro del subconjunto de entrenamiento de cada fold, mientras que el suavizado de etiquetas contribuyó a estabilizar el aprendizaje y reducir sobreconfianza en las predicciones.

Adicionalmente, se utilizó *gradient clipping* con norma máxima de 1.0 para evitar inestabilidad en la optimización, y la tasa de aprendizaje se programó mediante un esquema *Cosine Annealing* a lo largo del corto tramo de ajuste fino. La selección del mejor *checkpoint* en cada fold se realizó con base en el *macro-F1* de validación, métrica especialmente adecuada en este problema por su sensibilidad al equilibrio entre precisión y recall en las tres clases.

En conjunto, esta etapa puede describirse como un ajuste fino breve, estratificado y parcialmente descongelado, orientado a refinar las representaciones de alto nivel del modelo base sin modificar su estructura original ni sobreajustar sus parámetros más generales.

6.5.1.1. Resultados con Fine-Tuning.

El proceso de fine-tuning aplicado al modelo CSWin-Transformer + FPN permitió refinar de manera controlada los pesos del modelo base previamente entrenado. Al tratarse de un ajuste corto y focalizado sobre los módulos superiores y el *stage 4* del backbone, el objetivo principal fue mejorar la adaptación final al dominio OCT y optimizar la separabilidad entre clases manteniendo la estabilidad del modelo.



(a) Gráfica de exactitud aplicando Fine Tuning.

(b) Gráfica de pérdida aplicando Fine Tuning.

Figura 6.7: Curvas de entrenamiento con Fine-Tuning para el modelo CSWin-Transformer + FPN.

En la Figura 6.7 se observa la evolución de la exactitud y la pérdida durante el ajuste fino. En términos generales, las curvas muestran un entrenamiento estable a lo largo de los 5 folds, con mejoras rápidas desde las primeras épocas y sin oscilaciones abruptas. Esto es consistente con el hecho de que el modelo ya partía de un estado previamente optimizado y sólo requería un refinamiento sobre sus capas de más alto nivel semántico.

La curva de exactitud refleja una adaptación consistente durante el tramo de ajuste, mientras que la pérdida mantiene una tendencia controlada tanto en entrenamiento como en validación. En conjunto, estas curvas sugieren que el fine-tuning logró refinar el modelo sin introducir inestabilidad ni evidencia de sobreajuste severo, aun cuando la actualización del backbone se limitó únicamente al último bloque.

Métricas globales promedio.

Con el fin de resumir cuantitativamente el comportamiento general del modelo ajustado, en la Tabla 6.9 se presentan las métricas promedio obtenidas en los 5 folds sobre el conjunto de prueba externo.

Tabla 6.9: Métricas globales promedio del modelo CSWin-Transformer + FPN con Fine-Tuning en el conjunto de prueba externo.

Métrica	**Valor**
Accuracy	0.9401 ± 0.0043
Precision macro	0.9397 ± 0.0053
Recall macro	0.9374 ± 0.0041
F1 macro	0.9379 ± 0.0050
Precision weighted	0.9404 ± 0.0040
Recall weighted	0.9401 ± 0.0043
F1 weighted	0.9395 ± 0.0045
Macro-AUC	0.9753 ± 0.0055
Micro-AUC	0.9780 ± 0.0058

La exactitud promedio alcanzada fue de 0.9401 ± 0.0043, lo que indica que el modelo clasifica correctamente, en promedio, el 94.01 % de las imágenes del conjunto de prueba externo. La baja desviación estándar sugiere además un comportamiento estable entre folds, aun cuando el ajuste fino se realizó en un número reducido de épocas.

El *precision macro* de 0.9397 ± 0.0053 y el *recall macro* de 0.9374 ± 0.0041 muestran que el modelo mantiene un comportamiento equilibrado al considerar las tres clases con el mismo peso. La cercanía entre ambas métricas sugiere que el fine-tuning no introduce un sesgo importante hacia precisión o sensibilidad, sino que conserva un balance adecuado entre ambas dimensiones.

El *F1 macro* de 0.9379 ± 0.0050 resume este equilibrio entre precisión y recall a nivel interclase. Dado que esta métrica no favorece a la clase mayoritaria, su valor confirma que el modelo ajustado mantiene un rendimiento sólido en el problema multiclase. De manera complementaria, el *F1 weighted* de 0.9395 ± 0.0045 resulta muy cercano al F1 macro, lo que indica que el desbalance del conjunto no altera de forma relevante la interpretación global del desempeño.

Costo computacional y tiempos de ejecución.

Además del desempeño predictivo, resulta importante analizar el costo computacional asociado al ajuste fino y a la inferencia del modelo. La Tabla 6.10 resume los tiempos promedio registrados durante los 5 folds.

Tabla 6.10: Tiempos promedio de entrenamiento e inferencia del modelo CSWin-Transformer + FPN con Fine-Tuning en GPU A100.

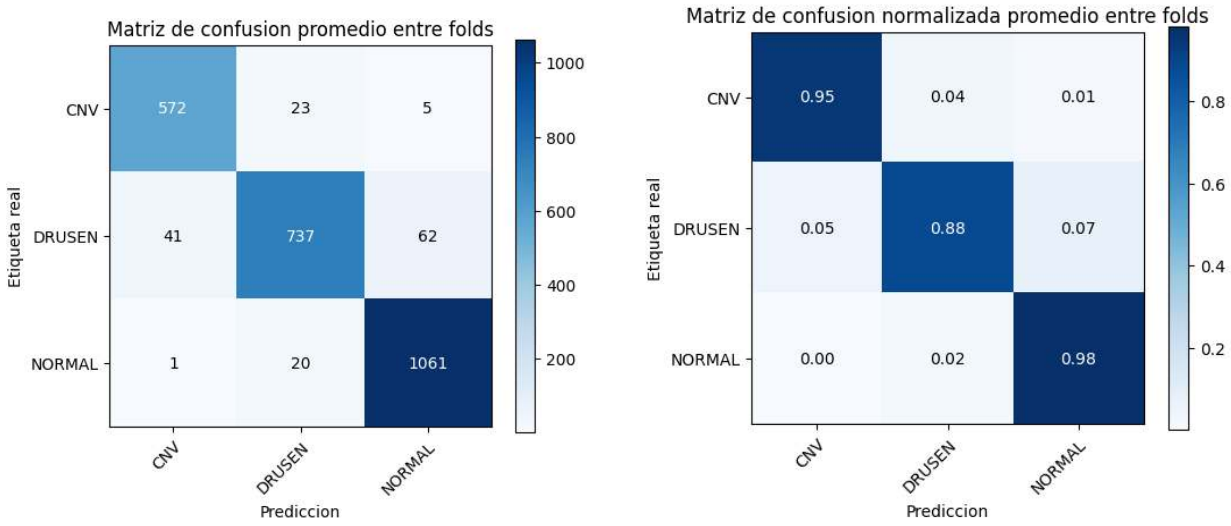
Métrica temporal	**Valor**
Tiempo total de entrenamiento por fold	222.5607 ± 3.1713 s
Tiempo total de validación por fold	43.2269 ± 3.2078 s
Tiempo total por fold	266.4067 ± 2.3308 s
Tiempo de inferencia en test	18.3025 ± 0.0169 s
Tiempo de inferencia por imagen	7.2600 ± 0.0067 ms
Imágenes procesadas por segundo	137.7411 ± 0.1275

En promedio, cada fold requirió 222.5607 ± 3.1713 segundos de entrenamiento y 43.2269 ± 3.2078 segundos de validación, para un tiempo total de 266.4067 ± 2.3308 segundos por fold. Esto equivale aproximadamente a 3.71 minutos de entrenamiento puro y 4.44 minutos de ejecución total por fold, lo cual confirma que se trató de un ajuste fino breve y computacionalmente mucho más ligero que el entrenamiento base.

En la etapa de inferencia sobre el conjunto de prueba externo, el modelo registró un tiempo promedio de 18.3025 ± 0.0169 segundos por evaluación completa, con un costo de 7.2600 ± 0.0067 ms por imagen y una velocidad aproximada de 137.7411 ± 0.1275 imágenes por segundo. Estos valores indican que el ajuste fino no comprometió la eficiencia de inferencia del modelo.

Estos tiempos deben interpretarse en el contexto del hardware utilizado, específicamente una GPU NVIDIA A100. La capacidad de cómputo paralelo y el soporte para operaciones tensoriales aceleradas de esta GPU permitieron ejecutar tanto el entrenamiento base como el ajuste fino en tiempos razonables, incluso bajo un protocolo de validación cruzada de 5 folds. En este caso, la A100 resultó especialmente útil para hacer viable una estrategia incremental de refinamiento sobre una arquitectura transformer con módulo multiescala.

Análisis de la matriz de confusión.



(a) Matriz de confusión aplicando Fine Tuning.

(b) Matriz de confusión normalizada.

La matriz de confusión promedio mostrada en la Figura 6.8a resume el comportamiento del modelo sobre el conjunto de prueba externo al promediar los resultados de los 5 folds. En términos de conteos promedio, la matriz indica aproximadamente 572.0 casos de CNV correctamente clasificados, 736.6 casos de DRUSEN correctamente identificados y 1061.4 imágenes NORMAL reconocidas de manera adecuada.

La matriz normalizada de la Figura 6.8b permite analizar con mayor claridad la sensibilidad promedio por clase. Los recalls promedio obtenidos fueron de 0.9533 ± 0.0129 para CNV, 0.8779 ± 0.0197 para DRUSEN y 0.9810 ± 0.0020 para NORMAL. Estos resultados muestran nuevamente que la clase NORMAL es la mejor recuperada por el modelo, mientras que DRUSEN continúa siendo la categoría de mayor dificultad relativa.

El patrón de error observado indica que la principal fuente de confusión se concentra en la clase DRUSEN. En promedio, esta clase presenta confusiones tanto hacia CNV como hacia NORMAL, aunque el mayor volumen de error se mantiene en la dirección DRUSEN $\rightarrow$ NORMAL/CNV. Este comportamiento sigue siendo clínicamente plausible, ya que ciertas manifestaciones estructurales de DRUSEN pueden compartir características con otras alteraciones maculares o con regiones de apariencia menos marcada.

Métricas promedio por clase.

Para profundizar en el análisis interclase, la Tabla 6.11 presenta las métricas promedio por clase obtenidas en el conjunto de prueba externo.

Tabla 6.11: Métricas promedio por clase del modelo CSWin-Transformer + FPN con Fine-Tuning en el conjunto de prueba externo.

Clase	**Precisión**	**Recall**	**F1-score**	**AUC**
CNV	0.9330 ± 0.0195	0.9533 ± 0.0129	0.9429 ± 0.0063	0.9788 ± 0.0053
DRUSEN	0.9457 ± 0.0080	0.8779 ± 0.0197	0.9104 ± 0.0082	0.9592 ± 0.0120
NORMAL	0.9405 ± 0.0047	0.9810 ± 0.0020	0.9603 ± 0.0025	0.9868 ± 0.0013

La clase **CNV** presenta un desempeño alto y equilibrado, con recall ligeramente superior a la precisión, lo que indica una buena capacidad de recuperación de casos verdaderos y una tasa contenida de errores de asignación. La clase **NORMAL** mantiene el recall más alto del modelo, lo que confirma que la mayoría de los casos normales son identificados correctamente.

En **DRUSEN**, la precisión permanece alta (0.9457 ± 0.0080), pero el recall se reduce a 0.8779 ± 0.0197, lo que indica que una parte de los casos reales continúa siendo confundida con otras categorías. Esta diferencia entre precisión y recall explica que el F1-score de DRUSEN (0.9104 ± 0.0082) sea el más bajo entre las tres clases, confirmando que esta categoría sigue siendo la más desafiante para la arquitectura aun después del ajuste fino.

Curvas ROC y capacidad discriminativa.

Las curvas ROC promedio bajo el esquema One-vs-Rest, mostradas en la Figura 6.9, permiten analizar la capacidad discriminativa del modelo ajustado de forma independiente para cada clase. Los valores promedio de AUC fueron de 0.9788 ± 0.0053 para CNV, 0.9592 ± 0.0120 para DRUSEN y 0.9868 ± 0.0013 para NORMAL.

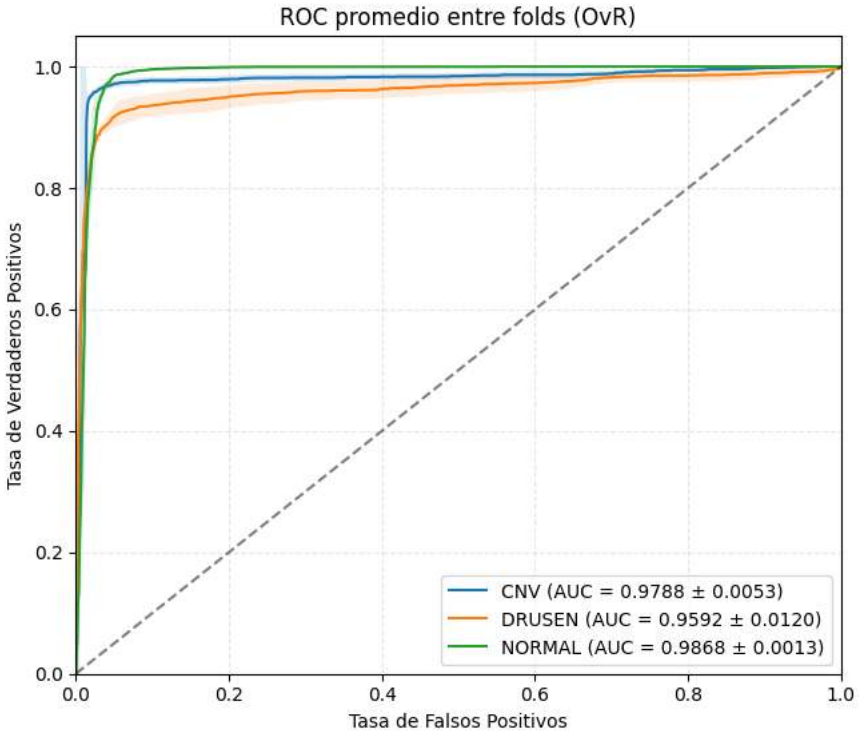


Figura 6.9: Curvas ROC de las 3 clases.

Estos resultados indican que el modelo mantiene una capacidad alta de separación entre clases al considerar distintos umbrales de decisión. En particular, la clase NORMAL presenta la mayor separabilidad global, mientras que DRUSEN vuelve a mostrar el valor más bajo, en concordancia con lo observado previamente en la matriz de confusión y en las métricas de recuperación.

A nivel global, el *macro-AUC* alcanzó 0.9753 ± 0.0055 y el *micro-AUC* fue de 0.9780 ± 0.0058. La cercanía entre ambas métricas indica que el modelo conserva una separabilidad alta y relativamente uniforme en el problema de clasificación multiclase.

Síntesis del desempeño.

En conjunto, el modelo CSWin-Transformer + FPN con Fine-Tuning demuestra:

- Convergencia estable durante el ajuste fino promedio de los 5 folds.
- Exactitud global de 0.9401 ± 0.0043.
- Precision macro de 0.9397 ± 0.0053, recall macro de 0.9374 ± 0.0041 y F1 macro de 0.9379 ± 0.0050.
- Weighted-F1 de 0.9395 ± 0.0045, consistente con el desempeño global del modelo.
- Macro-AUC de 0.9753 ± 0.0055 y Micro-AUC de 0.9780 ± 0.0058.
- Tiempo promedio de entrenamiento de 222.5607 ± 3.1713 s por fold e inferencia de 18.3025 ± 0.0169 s sobre el conjunto de prueba externo.

- Velocidad de inferencia de 137.7411 ± 0.1275 imágenes por segundo en GPU A100.
- Desempeño alto en las tres clases, con mayor dificultad relativa en la identificación de DRUSEN.

6.5.1.2. Comparación sin Fine-Tuning.

Con el propósito de cuantificar el efecto del ajuste fino, la Tabla 6.12 compara el desempeño promedio del modelo base CSWin-Transformer + FPN frente a su versión ajustada mediante Fine-Tuning.

Tabla 6.12: Comparación CSWin-Transformer + FPN antes y después de Fine-Tuning

Métrica	**FPN (Base)**	**FPN + FT**	**Δ**
Accuracy	0.9352 ± 0.0049	$\mathbf{0.9401 \pm 0.0043}$	+0.0049
Precision macro	0.9394 ± 0.0039	$\mathbf{0.9397 \pm 0.0053}$	+0.0003
Recall macro	0.9306 ± 0.0054	$\mathbf{0.9374 \pm 0.0041}$	+0.0068
Macro-F1	0.9336 ± 0.0045	$\mathbf{0.9379 \pm 0.0050}$	+0.0043
Weighted-F1	0.9344 ± 0.0051	$\mathbf{0.9395 \pm 0.0045}$	+0.0051
Macro-AUC (OvR)	0.9740 ± 0.0049	$\mathbf{0.9753 \pm 0.0055}$	+0.0013
Micro-AUC (OvR)	0.9766 ± 0.0052	$\mathbf{0.9780 \pm 0.0058}$	+0.0014

La Tabla 6.12 muestra que el fine-tuning produce una mejora global moderada pero consistente respecto al modelo base. La exactitud aumenta de 0.9352 ± 0.0049 a 0.9401 ± 0.0043, mientras que el *Macro-F1* se incrementa de 0.9336 ± 0.0045 a 0.9379 ± 0.0050. Estos cambios indican que el ajuste fino no sólo mejora el número total de clasificaciones correctas, sino también el equilibrio entre precisión y recall al considerar las tres clases con el mismo peso.

La mejora más clara se observa en el *recall macro*, que pasa de 0.9306 ± 0.0054 a 0.9374 ± 0.0041. Esto sugiere que el modelo ajustado recupera una mayor proporción de casos verdaderos en términos globales, lo que es especialmente relevante en un problema clínico donde reducir falsos negativos resulta importante. En paralelo, el *Macro-AUC* y el *Micro-AUC* también aumentan, lo que indica una ligera mejora en la separabilidad probabilística de las clases.

A nivel por clase, el impacto del ajuste fino no es uniforme. La mejora más relevante se observa en **DRUSEN**, cuyo recall aumenta de 0.8610 ± 0.0241 a 0.8779 ± 0.0197 y cuyo F1-score pasa de 0.9023 ± 0.0083 a 0.9104 ± 0.0082. También se aprecia una mejora en la clase **NORMAL**, particularmente en precisión y F1-score. En contraste, la clase **CNV** muestra un incremento en recall pero una ligera disminución en precisión, lo que hace que su F1-score permanezca prácticamente estable.

Desde una perspectiva representacional, estos resultados sugieren que el ajuste fino parcial sobre FPN, cabeza y *stage 4* del backbone permitió refinar las fronteras de decisión del modelo, especialmente en aquellas regiones del espacio de características donde DRUSEN presentaba mayor ambigüedad. En este sentido, el beneficio del fine-tuning no se limita a una mejora marginal en exactitud, sino que también se manifiesta en una recuperación más equilibrada de clases difíciles y en una calibración probabilística ligeramente superior.

En consecuencia, el impacto del Fine-Tuning sobre la arquitectura FPN puede considerarse positivo y consistente. Aunque las mejoras globales son moderadas, el ajuste fino permitió obtener

un modelo ligeramente más robusto, con mejor sensibilidad promedio y mejor comportamiento en la clase más desafiante del problema.

6.5.2. Impacto en CSWin+ASPP.

Estrategia en CSWin-Transformer + ASPP.

El ajuste fino del modelo CSWin-Transformer + ASPP se realizó como una segunda etapa de optimización incremental a partir de los mejores *checkpoints* obtenidos en el entrenamiento base para cada uno de los 5 folds. En esta fase no se introdujeron cambios en la arquitectura, por lo que se mantuvieron intactos el backbone CSWin-Transformer, el módulo ASPP y la cabeza clasificadora final.

El objetivo del fine-tuning fue refinar las representaciones aprendidas por el modelo base y mejorar la adaptación final al dominio OCT sin reentrenar de manera extensa toda la arquitectura. Para ello, el ajuste fino se ejecutó durante 3 épocas por fold, utilizando como punto de partida el mejor modelo base correspondiente a cada partición.

La estrategia de actualización de parámetros se organizó en dos fases. Durante la primera época, únicamente se optimizaron los parámetros del módulo ASPP y de la cabeza clasificadora, manteniendo completamente congelado el backbone. A partir de la segunda época, se habilitó el entrenamiento del *stage 4* del backbone, mientras que los niveles anteriores permanecieron congelados. Este esquema permitió conservar las representaciones más generales aprendidas por las etapas profundas iniciales del backbone y concentrar el ajuste en los niveles de mayor contenido semántico, que son los más relevantes para la decisión final de clasificación.

La optimización se realizó con AdamW y tasas de aprendizaje diferenciadas por bloque:

- Backbone CSWin-Transformer (*stage 4*): 5×10^{-6}
- ASPP: 3×10^{-5}
- Cabeza clasificadora: 5×10^{-5}

Esta configuración favorece un ajuste conservador sobre el backbone y una mayor plasticidad en los módulos superiores responsables de la integración contextual multiescala y la clasificación final.

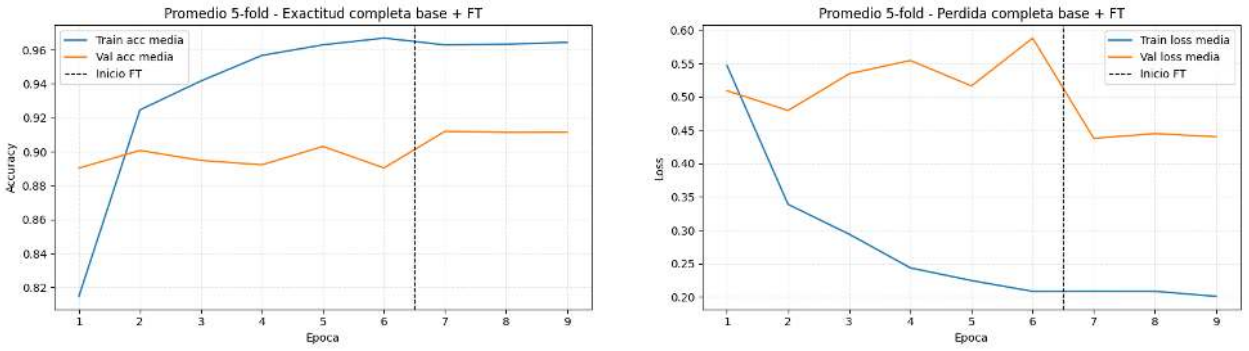
Como función de pérdida se empleó `CrossEntropyLoss` con pesos por clase calculados dinámicamente en cada fold y *label smoothing* de 0.02. La incorporación de pesos de clase permitió compensar diferencias de frecuencia entre categorías dentro del subconjunto de entrenamiento de cada fold, mientras que el suavizado de etiquetas contribuyó a estabilizar el aprendizaje y reducir sobreconfianza en las predicciones.

Adicionalmente, se utilizó *gradient clipping* con norma máxima de 1.0 para evitar inestabilidad en la optimización, y la tasa de aprendizaje se programó mediante un esquema *Cosine Annealing* a lo largo del corto tramo de ajuste fino. La selección del mejor *checkpoint* en cada fold se realizó con base en el *macro-F1* de validación, métrica especialmente pertinente en este problema por su sensibilidad al equilibrio entre precisión y recall en las tres clases.

En conjunto, esta etapa puede describirse como un ajuste fino breve, estratificado y parcialmente descongelado, orientado a refinar las representaciones de alto nivel del modelo base sin modificar la arquitectura original.

6.5.2.1. Resultados con Fine-Tuning.

El proceso de fine-tuning aplicado al modelo CSWin-Transformer + ASPP permitió refinar de manera controlada los pesos del modelo base previamente entrenado. Dado que el ajuste se concentró en el módulo ASPP, la cabeza clasificadora y el *stage 4* del backbone, el propósito principal fue optimizar la separabilidad entre clases y mejorar la adaptación al dominio OCT manteniendo la estabilidad del modelo.



(a) Gráfica de exactitud aplicando Fine Tuning.

(b) Gráfica de pérdida aplicando Fine Tuning.

Figura 6.10: Curvas de entrenamiento con Fine-Tuning para el modelo CSWin-Transformer+ASPP.

En la Figura 6.10 se observa la evolución de la exactitud y la pérdida durante el ajuste fino. En términos generales, las curvas muestran un entrenamiento estable a lo largo de los 5 folds, con mejoras rápidas desde las primeras épocas y sin oscilaciones abruptas. Este comportamiento es consistente con el hecho de que el modelo ya partía de un estado previamente optimizado y sólo requería un refinamiento localizado sobre los módulos de mayor nivel semántico.

La evolución de la pérdida también refleja un comportamiento controlado tanto en entrenamiento como en validación. En conjunto, estas curvas sugieren que el fine-tuning logró refinar el modelo sin introducir inestabilidad ni evidencia de sobreajuste severo, aun cuando el ajuste se realizó en un tramo corto de sólo 3 épocas por fold.

Métricas globales promedio.

Con el fin de resumir cuantitativamente el comportamiento general del modelo ajustado, en la Tabla 6.13 se presentan las métricas promedio obtenidas en los 5 folds sobre el conjunto de prueba externo.

Tabla 6.13: Métricas globales promedio del modelo CSWin-Transformer + ASPP con Fine-Tuning en el conjunto de prueba externo.

Métrica	**Valor**
Accuracy	0.9436 ± 0.0060
Precision macro	0.9450 ± 0.0041
Recall macro	0.9415 ± 0.0077
F1 macro	0.9426 ± 0.0061
Precision weighted	0.9440 ± 0.0056
Recall weighted	0.9436 ± 0.0060
F1 weighted	0.9431 ± 0.0061
Macro-AUC	0.9797 ± 0.0019
Micro-AUC	0.9811 ± 0.0026

La exactitud promedio alcanzada fue de 0.9436 ± 0.0060, lo que indica que el modelo clasifica correctamente, en promedio, el 94.36 % de las imágenes del conjunto de prueba externo. La desviación estándar relativamente baja muestra además un comportamiento estable entre folds, aun cuando el ajuste fino se realizó en un número reducido de épocas.

El *precision macro* de 0.9450 ± 0.0041 y el *recall macro* de 0.9415 ± 0.0077 muestran que el modelo mantiene un comportamiento equilibrado al considerar las tres clases con el mismo peso. La cercanía entre ambas métricas sugiere que el fine-tuning no introduce un sesgo relevante hacia precisión o sensibilidad, sino que conserva un balance adecuado entre ambas dimensiones.

El *F1 macro* de 0.9426 ± 0.0061 resume este equilibrio entre precisión y recall a nivel interclase. Al no favorecer a la clase mayoritaria, esta métrica confirma que el modelo ajustado conserva un desempeño sólido en el problema multiclase. De manera complementaria, el *F1 weighted* de 0.9431 ± 0.0061 resulta muy cercano al F1 macro, lo que indica que el desbalance del conjunto no altera de manera importante la interpretación global del rendimiento.

Costo computacional y tiempos de ejecución.

Además del desempeño predictivo, resulta importante analizar el costo computacional asociado al ajuste fino y a la inferencia del modelo. La Tabla 6.14 resume los tiempos promedio registrados durante los 5 folds.

Tabla 6.14: Tiempos promedio de entrenamiento e inferencia del modelo CSWin-Transformer + ASPP con Fine-Tuning en GPU A100.

Métrica temporal	**Valor**
Tiempo total de entrenamiento por fold	217.6479 ± 4.7842 s
Tiempo total de validación por fold	40.9588 ± 2.9226 s
Tiempo total por fold	259.1477 ± 3.2137 s
Tiempo de inferencia en test	17.3657 ± 0.0379 s
Tiempo de inferencia por imagen	6.8884 ± 0.0150 ms
Imágenes procesadas por segundo	145.1715 ± 0.3172

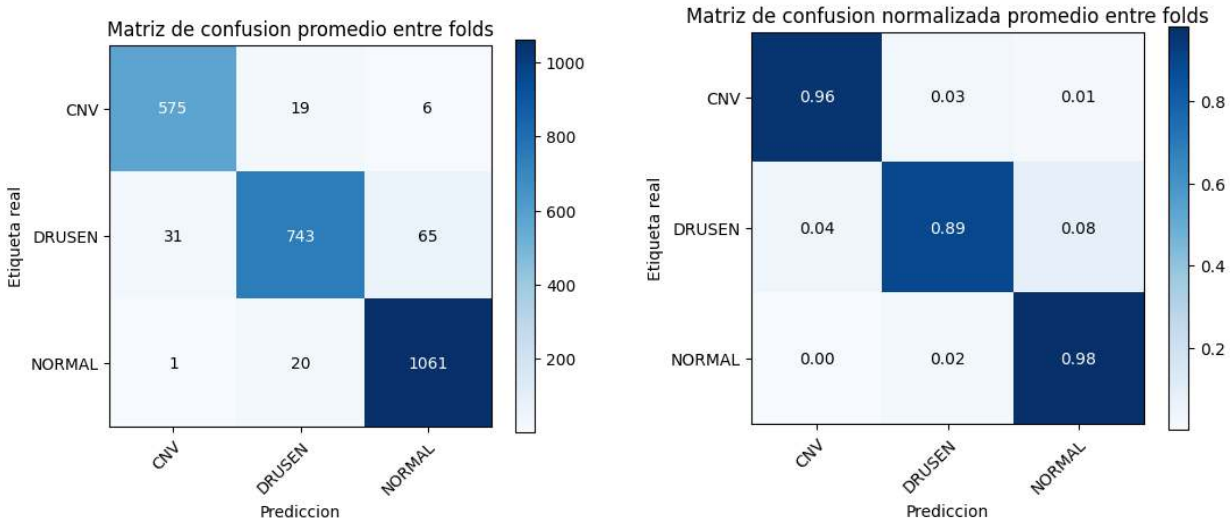
En promedio, cada fold requirió 217.6479 ± 4.7842 segundos de entrenamiento y 40.9588 ± 2.9226 segundos de validación, para un tiempo total de 259.1477 ± 3.2137 segundos por fold. Esto equivale

aproximadamente a 3.63 minutos de entrenamiento puro y 4.32 minutos de ejecución total por fold, lo cual confirma que se trató de un ajuste fino breve y computacionalmente mucho más ligero que el entrenamiento base.

En la etapa de inferencia sobre el conjunto de prueba externo, el modelo registró un tiempo promedio de 17.3657 ± 0.0379 segundos por evaluación completa, con un costo de 6.8884 ± 0.0150 ms por imagen y una velocidad aproximada de 145.1715 ± 0.3172 imágenes por segundo. Estos valores indican que el ajuste fino no comprometió la eficiencia de inferencia del modelo.

Estos tiempos deben interpretarse en el contexto del hardware utilizado, específicamente una GPU NVIDIA A100. La capacidad de cómputo paralelo y el soporte para operaciones tensoriales aceleradas de esta GPU permitieron ejecutar tanto el entrenamiento base como el ajuste fino en tiempos razonables, incluso bajo un protocolo de validación cruzada de 5 folds. En este caso, la A100 permitió que una estrategia incremental de refinamiento sobre una arquitectura transformer con módulo ASPP siguiera siendo computacionalmente viable.

Análisis de la matriz de confusión.



(a) Matriz de confusión aplicando Fine Tuning.

(b) Matriz de confusión normalizada.

La matriz de confusión promedio mostrada en la Figura 6.11a resume el comportamiento del modelo sobre el conjunto de prueba externo al promediar los resultados de los 5 folds. En términos de conteos promedio, la matriz indica aproximadamente 575.2 casos de CNV correctamente clasificados, 742.8 casos de DRUSEN correctamente identificados y 1060.8 imágenes NORMAL reconocidas de manera adecuada.

La matriz normalizada de la Figura 6.11b permite analizar con mayor claridad la sensibilidad promedio por clase. Los recalls promedio obtenidos fueron de 0.9587 ± 0.0154 para CNV, 0.8853 ± 0.0135 para DRUSEN y 0.9804 ± 0.0050 para NORMAL. Estos resultados muestran nuevamente que la clase NORMAL es la mejor recuperada por el modelo, mientras que DRUSEN continúa siendo la categoría de mayor dificultad relativa.

El patrón de error observado indica que la principal fuente de confusión se concentra en la clase DRUSEN. En promedio, esta clase presenta confusiones tanto hacia CNV como hacia NORMAL,

aunque el mayor volumen de error se mantiene hacia las restantes categorías diagnósticas. Este comportamiento sigue siendo clínicamente razonable, ya que ciertas manifestaciones estructurales de DRUSEN pueden compartir rasgos morfológicos con otras alteraciones maculares o con regiones de apariencia menos marcada.

Métricas promedio por clase.

Para profundizar en el análisis interclase, la Tabla 6.15 presenta las métricas promedio por clase obtenidas en el conjunto de prueba externo.

Tabla 6.15: Métricas promedio por clase del modelo CSWin-Transformer + ASPP con Fine-Tuning en el conjunto de prueba externo.

Clase	**Precisión**	**Recall**	**F1-score**	**AUC**
CNV	0.9475 ± 0.0117	0.9587 ± 0.0154	0.9529 ± 0.0049	0.9857 ± 0.0047
DRUSEN	0.9501 ± 0.0075	0.8853 ± 0.0135	0.9166 ± 0.0098	0.9646 ± 0.0042
NORMAL	0.9373 ± 0.0124	0.9804 ± 0.0050	0.9583 ± 0.0046	0.9879 ± 0.0016

La clase **CNV** presenta un desempeño alto y equilibrado, con precisión de 0.9475 ± 0.0117, recall de 0.9587 ± 0.0154 y F1-score de 0.9529 ± 0.0049. Esto indica que el modelo identifica correctamente una gran proporción de los casos reales de CNV y mantiene una tasa contenida de errores de asignación.

En **DRUSEN**, la precisión permanece alta (0.9501 ± 0.0075), lo que sugiere que cuando el modelo predice esta clase suele hacerlo correctamente. Sin embargo, el recall disminuye a 0.8853 ± 0.0135, lo que indica que una fracción de los casos reales de DRUSEN continúa siendo confundida con otras categorías. Esta diferencia entre precisión y recall explica que el F1-score de DRUSEN (0.9166 ± 0.0098) sea el más bajo entre las tres clases, confirmando que esta categoría sigue siendo la más desafiante para la arquitectura aun después del ajuste fino.

La clase **NORMAL** alcanza el recall más alto del modelo (0.9804 ± 0.0050), lo que indica que la gran mayoría de los casos normales son detectados correctamente. Aunque su precisión (0.9373 ± 0.0124) es menor que la observada en CNV y DRUSEN, su F1-score se mantiene elevado (0.9583 ± 0.0046). Esto sugiere que la arquitectura conserva una muy buena capacidad para identificar normalidad, aunque en algunos casos todavía asigna esta categoría a imágenes patológicas con características menos marcadas.

Curvas ROC y capacidad discriminativa.

Las curvas ROC promedio bajo el esquema One-vs-Rest, mostradas en la Figura 6.12, permiten analizar la capacidad discriminativa del modelo ajustado de forma independiente para cada clase. Los valores promedio de AUC fueron de 0.9857 ± 0.0047 para CNV, 0.9646 ± 0.0042 para DRUSEN y 0.9879 ± 0.0016 para NORMAL.

Estos resultados indican que el modelo mantiene una alta capacidad de separación entre clases al considerar distintos umbrales de decisión. En particular, la clase NORMAL presenta la mayor separabilidad global, seguida muy de cerca por CNV, mientras que DRUSEN vuelve a mostrar el valor más bajo, en concordancia con lo observado previamente en la matriz de confusión y en las métricas de recuperación.

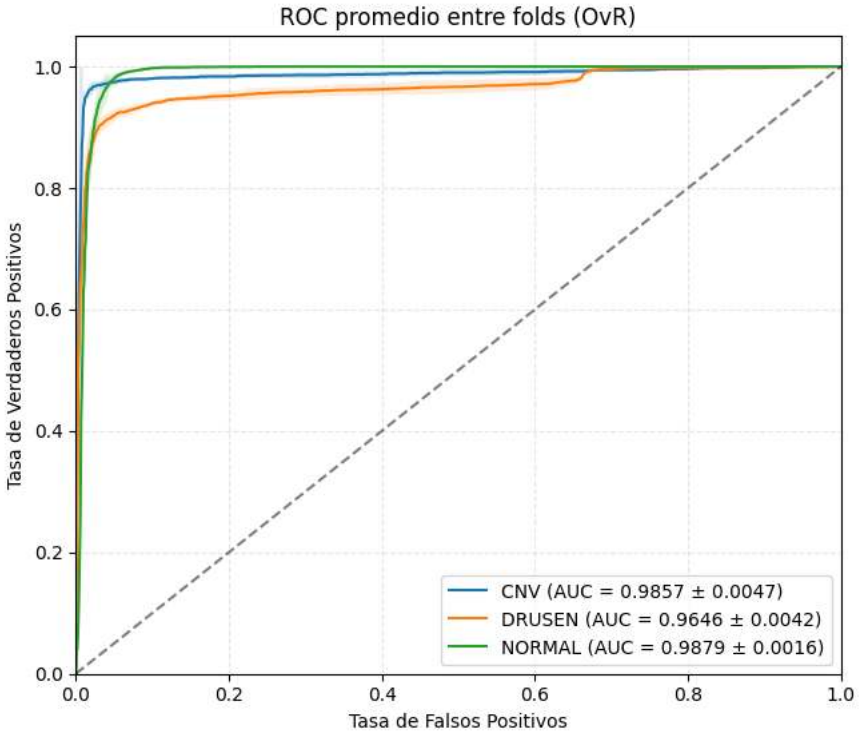


Figura 6.12: Curvas ROC de las 3 clases.

A nivel global, el *macro-AUC* alcanzó 0.9797 ± 0.0019 y el *micro-AUC* fue de 0.9811 ± 0.0026. La cercanía entre ambas métricas indica que el modelo conserva una separabilidad alta y relativamente uniforme en el problema de clasificación multiclase.

Síntesis del desempeño.

En conjunto, el modelo CSWin-Transformer + ASPP con Fine-Tuning demuestra:

- Convergencia estable durante el ajuste fino promedio de los 5 folds.
- Exactitud global de 0.9436 ± 0.0060.
- Precision macro de 0.9450 ± 0.0041, recall macro de 0.9415 ± 0.0077 y F1 macro de 0.9426 ± 0.0061.
- Weighted-F1 de 0.9431 ± 0.0061, consistente con el desempeño global del modelo.
- Macro-AUC de 0.9797 ± 0.0019 y Micro-AUC de 0.9811 ± 0.0026.
- Tiempo promedio de entrenamiento de 217.6479 ± 4.7842 s por fold e inferencia de 17.3657 ± 0.0379 s sobre el conjunto de prueba externo.
- Velocidad de inferencia de 145.1715 ± 0.3172 imágenes por segundo en GPU A100.
- Desempeño alto en las tres clases, con mayor dificultad relativa en la identificación de DRUSEN.

6.5.2.2. Comparación sin Fine-Tuning.

Con el propósito de cuantificar el efecto del ajuste fino, la Tabla 6.16 compara el desempeño promedio del modelo base CSWin-Transformer + ASPP frente a su versión ajustada mediante Fine-Tuning.

Tabla 6.16: Comparación CSWin-Transformer + ASPP antes y después de Fine-Tuning

Métrica	**ASPP (Base)**	**ASPP + FT**	**Δ**
Accuracy	$\mathbf{0.9440 \pm 0.0073}$	0.9436 ± 0.0060	-0.0004
Precision macro	$\mathbf{0.9470 \pm 0.0049}$	0.9450 ± 0.0041	-0.0020
Recall macro	0.9408 ± 0.0090	$\mathbf{0.9415 \pm 0.0077}$	$+0.0007$
Macro-F1	$\mathbf{0.9433 \pm 0.0069}$	0.9426 ± 0.0061	-0.0007
Weighted-F1	$\mathbf{0.9436 \pm 0.0072}$	0.9431 ± 0.0061	-0.0005
Macro-AUC (OvR)	$\mathbf{0.9799 \pm 0.0014}$	0.9797 ± 0.0019	-0.0002
Micro-AUC (OvR)	$\mathbf{0.9811 \pm 0.0011}$	0.9811 ± 0.0026	≈ 0.0000

La Tabla 6.16 muestra que el fine-tuning produce un impacto muy limitado sobre el modelo base CSWin-Transformer + ASPP. La exactitud pasa de 0.9440 ± 0.0073 a 0.9436 ± 0.0060, mientras que el *Macro-F1* cambia de 0.9433 ± 0.0069 a 0.9426 ± 0.0061. Estas diferencias son pequeñas y sugieren que la arquitectura base ya se encontraba muy próxima a su punto de convergencia óptimo en términos de clasificación global.

En términos de sensibilidad promedio, el *recall macro* muestra una ligera mejora al pasar de 0.9408 ± 0.0090 a 0.9415 ± 0.0077, lo que indica una recuperación apenas superior de casos verdaderos al considerar las tres clases con el mismo peso. No obstante, esta ganancia se acompaña de leves reducciones en precisión macro, F1 macro y weighted-F1, por lo que el balance global del modelo permanece esencialmente estable.

La capacidad discriminativa también se mantiene prácticamente sin cambios. El *Macro-AUC* pasa de 0.9799 ± 0.0014 a 0.9797 ± 0.0019, mientras que el *Micro-AUC* se mantiene en 0.9811 con diferencias despreciables. Esto sugiere que el ajuste fino no modificó de forma sustancial la separabilidad probabilística de las clases, lo cual refuerza la idea de que el modelo base ASPP ya había aprendido representaciones altamente adecuadas para el problema.

A nivel por clase, el comportamiento también es muy estable. La clase **CNV** mejora ligeramente en recall, mientras que **DRUSEN** reduce marginalmente su recall respecto al modelo base (0.8894 ± 0.0080 frente a 0.8853 ± 0.0135) y mantiene un F1-score muy similar (0.9176 ± 0.0104 frente a 0.9166 ± 0.0098). En la clase **NORMAL**, las métricas permanecen prácticamente invariantes. En conjunto, estos resultados indican que el ajuste fino no alteró de manera significativa la estructura de errores del modelo.

Desde una perspectiva representacional, esto sugiere que el módulo ASPP ya estaba explotando de forma eficaz el contexto multiescala y el campo receptivo ampliado desde el entrenamiento base, por lo que el margen de mejora adicional mediante un ajuste fino corto y parcial fue reducido. En otras palabras, el modelo base ASPP ya mostraba una adaptación muy sólida al dominio OCT, y el fine-tuning actuó más como un refinamiento conservador que como una etapa de mejora sustancial.

En consecuencia, el impacto del Fine-Tuning sobre la arquitectura ASPP puede considerarse neutro o marginal. Aunque el ajuste fino no deteriora de forma importante el desempeño, tampoco aporta una mejora clara frente al modelo base, lo que sugiere que la variante CSWin-Transformer

+ ASPP alcanza su mejor compromiso entre precisión, estabilidad y separabilidad desde la fase base de entrenamiento.

6.6. Comparación: CSWin-Transformer + FPN Fine-Tuning vs CSWin-Transformer + ASPP Fine-Tuning.

Dado que ambas variantes fueron sometidas al mismo protocolo de ajuste fino, resulta posible realizar una comparación directa entre CSWin-Transformer + FPN con Fine-Tuning y CSWin-Transformer + ASPP con Fine-Tuning. En ambos casos se empleó un esquema de ajuste corto de 3 épocas por fold, con entrenamiento en dos fases, apertura del *stage 4* del backbone a partir de la segunda época, optimización con AdamW, *label smoothing*, pesos de clase y selección del mejor *checkpoint* mediante *macro-F1* de validación. Por tanto, las diferencias observadas en el desempeño pueden atribuirse principalmente al tipo de módulo multiescala incorporado, y no a variaciones en la estrategia de ajuste.

Tabla 6.17: Comparación de desempeño entre CSWin-Transformer + FPN Fine-Tuning y CSWin-Transformer + ASPP Fine-Tuning

Métrica	**CSWin + FPN + FT**	**CSWin + ASPP + FT**
Accuracy	0.9401 ± 0.0043	$\mathbf{0.9436 \pm 0.0060}$
Precision macro	0.9397 ± 0.0053	$\mathbf{0.9450 \pm 0.0041}$
Recall macro	0.9374 ± 0.0041	$\mathbf{0.9415 \pm 0.0077}$
Macro-F1	0.9379 ± 0.0050	$\mathbf{0.9426 \pm 0.0061}$
Weighted-F1	0.9395 ± 0.0045	$\mathbf{0.9431 \pm 0.0061}$
Macro-AUC (OvR)	0.9753 ± 0.0055	$\mathbf{0.9797 \pm 0.0019}$
Micro-AUC (OvR)	0.9780 ± 0.0058	$\mathbf{0.9811 \pm 0.0026}$

La Tabla 6.17 muestra que, tras aplicar Fine-Tuning, la variante CSWin-Transformer + ASPP mantiene una ventaja consistente sobre CSWin-Transformer + FPN en todas las métricas globales evaluadas. En términos de exactitud, ASPP + FT alcanza 0.9436 ± 0.0060, superando el valor de 0.9401 ± 0.0043 obtenido por FPN + FT. Esta diferencia indica que, aun después del ajuste fino, ASPP conserva una ligera superioridad en la proporción global de imágenes correctamente clasificadas.

La misma tendencia se observa en las métricas de balance interclase. El *Macro-F1* aumenta de 0.9379 ± 0.0050 con FPN + FT a 0.9426 ± 0.0061 con ASPP + FT, mientras que el *Weighted-F1* pasa de 0.9395 ± 0.0045 a 0.9431 ± 0.0061. Dado que el F1-score combina precisión y recall, estos resultados sugieren que ASPP sigue ofreciendo una respuesta más equilibrada entre las tres clases incluso después del ajuste fino.

La capacidad discriminativa también favorece a la variante con ASPP. El *Macro-AUC* se incrementa de 0.9753 ± 0.0055 a 0.9797 ± 0.0019, mientras que el *Micro-AUC* mejora de 0.9780 ± 0.0058 a 0.9811 ± 0.0026. Además de presentar valores más altos, ASPP + FT mantiene desviaciones estándar más bajas en las métricas AUC, lo que sugiere una separabilidad más estable entre folds.

Para complementar la comparación de desempeño, la Tabla 6.18 presenta los tiempos promedio

de entrenamiento e inferencia registrados para ambas arquitecturas ajustadas en GPU NVIDIA A100.

Tabla 6.18: Comparación de tiempos de ejecución entre CSWin-Transformer + FPN Fine-Tuning y CSWin-Transformer + ASPP Fine-Tuning en GPU A100

Métrica temporal	**CSWin + FPN + FT**	**CSWin + ASPP + FT**
Tiempo total de entrenamiento por fold	222.5607 ± 3.1713 s	$\mathbf{217.6479 \pm 4.7842}$ s
Tiempo total de validación por fold	43.2269 ± 3.2078 s	$\mathbf{40.9588 \pm 2.9226}$ s
Tiempo total por fold	266.4067 ± 2.3308 s	$\mathbf{259.1477 \pm 3.2137}$ s
Tiempo de inferencia en test	18.3025 ± 0.0169 s	$\mathbf{17.3657 \pm 0.0379}$ s
Tiempo de inferencia por imagen	7.2600 ± 0.0067 ms	$\mathbf{6.8884 \pm 0.0150}$ ms
Imágenes procesadas por segundo	137.7411 ± 0.1275	$\mathbf{145.1715 \pm 0.3172}$

Desde el punto de vista computacional, ASPP + FT también presenta una ligera ventaja sobre FPN + FT. El tiempo total de entrenamiento por fold disminuye de 222.5607 ± 3.1713 s a 217.6479 ± 4.7842 s, mientras que la inferencia sobre el conjunto de prueba externo también resulta más rápida, pasando de 18.3025 ± 0.0169 s a 17.3657 ± 0.0379 s. En términos de costo por imagen, ASPP + FT reduce el tiempo de inferencia de 7.2600 ± 0.0067 ms a 6.8884 ± 0.0150 ms y aumenta la velocidad de procesamiento de 137.7411 ± 0.1275 a 145.1715 ± 0.3172 imágenes por segundo.

Estos resultados adquieren mayor relevancia al considerar que ambas variantes fueron ajustadas bajo el mismo protocolo y sobre la misma infraestructura de hardware, específicamente una GPU NVIDIA A100. En consecuencia, la diferencia observada no depende de condiciones experimentales distintas, sino del modo en que cada arquitectura aprovecha el refinamiento del ajuste fino. En este sentido, ASPP + FT no sólo mantiene mejores métricas predictivas, sino también una relación ligeramente más favorable entre desempeño y costo computacional.

Si se analizan las diferencias arquitectónicas, ambas configuraciones parten del backbone CSWin-Transformer y comparten el mismo esquema de fine-tuning. Sin embargo, la forma en que cada variante explota las representaciones del backbone sigue siendo distinta. La FPN continúa integrando información de diferentes escalas mediante una fusión jerárquica *top-down*, mientras que ASPP aplica convoluciones dilatadas en paralelo sobre una representación de alto nivel, ampliando el campo receptivo efectivo sin perder resolución espacial.

En el contexto del ajuste fino, esta diferencia resulta relevante porque el refinamiento se concentra precisamente sobre los módulos superiores y el *stage 4* del backbone. En FPN + FT, el ajuste permite mejorar la integración entre niveles jerárquicos, lo que se traduce en una mejora moderada respecto al modelo base. En ASPP + FT, en cambio, el módulo ya partía de una representación muy fuerte desde la fase base, por lo que el margen de mejora adicional es menor. Aun así, la variante ASPP ajustada conserva mejores resultados absolutos que FPN ajustado.

Este comportamiento se refleja de manera clara en la clase DRUSEN, que continúa siendo la categoría más difícil en ambas arquitecturas. En FPN + FT, DRUSEN alcanza un recall de 0.8779 ± 0.0197 y un F1-score de 0.9104 ± 0.0082. En ASPP + FT, estos valores mejoran a 0.8853 ± 0.0135 y 0.9166 ± 0.0098, respectivamente. Asimismo, el AUC de DRUSEN aumenta de 0.9592 ± 0.0120 a 0.9646 ± 0.0042. Estos resultados sugieren que ASPP conserva una ventaja en la modelación de la variabilidad morfológica asociada a esta clase, incluso después del ajuste fino.

A nivel global, la comparación también permite extraer una conclusión importante sobre el

efecto del Fine-Tuning. Mientras que FPN sí obtiene una mejora moderada tras el ajuste fino respecto a su modelo base, ASPP prácticamente mantiene el mismo nivel de desempeño que ya presentaba en su versión base. Esto implica que, aunque ambos modelos fueron sometidos al mismo protocolo de refinamiento, ASPP llega al fine-tuning desde una posición inicial más fuerte y sigue conservando la mejor combinación entre precisión diagnóstica, capacidad discriminativa y eficiencia computacional.

En síntesis, tras aplicar el mismo esquema de ajuste fino a ambas arquitecturas, CSWin-Transformer + ASPP continúa posicionándose por encima de CSWin-Transformer + FPN al ofrecer:

- mejores métricas globales de clasificación,
- mayor capacidad discriminativa,
- mejor recuperación de la clase DRUSEN,
- y una inferencia ligeramente más eficiente en GPU A100.

Por ello, dentro del conjunto de variantes ajustadas mediante Fine-Tuning, CSWin-Transformer + ASPP se mantiene como la opción con mejor compromiso entre rendimiento predictivo, estabilidad entre folds y costo computacional.

6.7. Comparación general.

Con el fin de integrar los resultados obtenidos a lo largo del estudio, la Tabla 6.19 resume el desempeño global de las cuatro configuraciones evaluadas: CSWin-Transformer + FPN, CSWin-Transformer + FPN con Fine-Tuning, CSWin-Transformer + ASPP y CSWin-Transformer + ASPP con Fine-Tuning. Todas las métricas corresponden al promedio y desviación estándar obtenidos bajo el protocolo de validación cruzada de 5 folds y evaluación sobre el conjunto de prueba externo.

Tabla 6.19: Tabla global comparativa de resultados: CSWin-Transformer con FPN y ASPP, antes y después de Fine-Tuning

Métrica	**FPN Base**	**FPN + FT**	**ASPP Base**	**ASPP + FT**
Accuracy	0.9352 ± 0.0049	0.9401 ± 0.0043	$\mathbf{0.9440 \pm 0.0073}$	0.9436 ± 0.0060
Precision macro	0.9394 ± 0.0039	0.9397 ± 0.0053	$\mathbf{0.9470 \pm 0.0049}$	0.9450 ± 0.0041
Recall macro	0.9306 ± 0.0054	0.9374 ± 0.0041	0.9408 ± 0.0090	$\mathbf{0.9415 \pm 0.0077}$
Macro-F1	0.9336 ± 0.0045	0.9379 ± 0.0050	$\mathbf{0.9433 \pm 0.0069}$	0.9426 ± 0.0061
Weighted-F1	0.9344 ± 0.0051	0.9395 ± 0.0045	$\mathbf{0.9436 \pm 0.0072}$	0.9431 ± 0.0061
Macro-AUC (OvR)	0.9740 ± 0.0049	0.9753 ± 0.0055	$\mathbf{0.9799 \pm 0.0014}$	0.9797 ± 0.0019
Micro-AUC (OvR)	0.9766 ± 0.0052	0.9780 ± 0.0058	$\mathbf{0.9811 \pm 0.0011}$	$\mathbf{0.9811 \pm 0.0026}$

La Tabla 6.19 muestra que la arquitectura CSWin-Transformer + ASPP en su versión base alcanza el mejor desempeño global en la mayoría de las métricas evaluadas. En particular, presenta la mayor exactitud (0.9440 ± 0.0073), el mejor *Macro-F1* (0.9433 ± 0.0069), el mayor *Weighted-F1*

(0.9436 ± 0.0072) y el valor más alto de *Macro-AUC* (0.9799 ± 0.0014). Estos resultados indican que la variante ASPP base ofrece la combinación más favorable entre clasificación correcta, equilibrio interclase y capacidad discriminativa.

En contraste, el modelo CSWin-Transformer + FPN en su versión base presenta el desempeño global más bajo de las cuatro configuraciones, con una exactitud de 0.9352 ± 0.0049 y un *Macro-F1* de 0.9336 ± 0.0045. No obstante, al aplicar Fine-Tuning, esta variante muestra una mejora consistente en todas las métricas principales: la exactitud aumenta a 0.9401 ± 0.0043, el *Macro-F1* sube a 0.9379 ± 0.0050 y el *Macro-AUC* se incrementa a 0.9753 ± 0.0055. Esto indica que el ajuste fino sí tuvo un efecto positivo sobre la variante con FPN, particularmente en términos de recuperación global de casos y refinamiento de la separabilidad entre clases.

Por su parte, el comportamiento de CSWin-Transformer + ASPP frente al Fine-Tuning fue diferente. Aunque la versión ajustada mantiene un desempeño muy alto, no supera de forma clara a la variante base. La exactitud pasa de 0.9440 ± 0.0073 a 0.9436 ± 0.0060, el *Macro-F1* cambia de 0.9433 ± 0.0069 a 0.9426 ± 0.0061 y el *Macro-AUC* desciende ligeramente de 0.9799 ± 0.0014 a 0.9797 ± 0.0019. Aunque estas diferencias son pequeñas, sugieren que el modelo base ASPP ya se encontraba muy cercano a su punto de convergencia óptimo, por lo que el ajuste fino no aportó una mejora sustancial adicional.

Además del desempeño predictivo, resulta relevante considerar el costo computacional asociado a cada configuración. La Tabla 6.20 resume los tiempos promedio de entrenamiento e inferencia de las cuatro variantes en GPU NVIDIA A100.

Tabla 6.20: Comparación global de tiempos de ejecución entre las cuatro variantes evaluadas en GPU A100

Métrica temporal	**FPN Base**	**FPN + FT**	**ASPP Base**	**ASPP + FT**
Tiempo total de entrenamiento por fold	1179.0331 ± 14.9562 s	222.5607 ± 3.1713 s	1169.9877 ± 16.9603 s	$\mathbf{217.6479 \pm 4.7842}$ s
Tiempo total de validación por fold	88.1436 ± 6.2696 s	43.2269 ± 3.2078 s	83.1835 ± 5.9078 s	$\mathbf{40.9588 \pm 2.9226}$ s
Tiempo total por fold	1268.0447 ± 9.7730 s	266.4067 ± 2.3308 s	1256.2208 ± 15.1983 s	$\mathbf{259.1477 \pm 3.2137}$ s
Tiempo de inferencia en test	18.7601 ± 0.6371 s	18.3025 ± 0.0169 s	17.4758 ± 0.0997 s	$\mathbf{17.3657 \pm 0.0379}$ s
Tiempo de inferencia por imagen	0.0074 ± 0.0003 s	0.0073 ± 0.0000 s	0.0069 ± 0.0000 s	$\mathbf{0.0069 \pm 0.0000}$ s
Imágenes procesadas por segundo	134.4995 ± 4.3718	137.7411 ± 0.1275	144.2607 ± 0.8209	$\mathbf{145.1715 \pm 0.3172}$

Desde el punto de vista computacional, las variantes con ASPP también muestran una ligera ventaja sobre las variantes con FPN. En particular, ASPP base presenta mejor velocidad de inferencia que FPN base, y ASPP + FT registra el menor tiempo de inferencia por imagen y la mayor tasa de procesamiento en imágenes por segundo. Asimismo, los modelos ajustados mediante Fine-Tuning presentan tiempos de entrenamiento muy inferiores a sus versiones base, lo cual es consistente con el carácter breve y parcialmente descongelado de esta segunda etapa de optimización.

Estos resultados deben interpretarse en el contexto del hardware utilizado, específicamente una GPU NVIDIA A100. La elevada capacidad de cómputo paralelo y el soporte eficiente para operaciones tensoriales aceleradas permitieron ejecutar un protocolo de validación cruzada de 5 folds con arquitecturas basadas en transformadores y módulos multiescala en tiempos razonables. Sin embargo, aun bajo las mismas condiciones de hardware, ASPP demostró una relación ligeramente más favorable entre desempeño predictivo y eficiencia computacional.

En términos comparativos, pueden establecerse las siguientes observaciones:

- La mejor configuración global corresponde a **CSWin-Transformer + ASPP en su versión base**.

- El **Fine-Tuning resulta beneficioso principalmente para la variante con FPN**, donde produce mejoras moderadas pero consistentes en exactitud, recall macro, F1 y AUC.
- En la variante **ASPP**, el Fine-Tuning no aporta una mejora clara respecto al modelo base, lo que sugiere que la arquitectura ya logra una representación altamente adecuada desde la fase inicial de entrenamiento.
- Las arquitecturas con **ASPP muestran una ligera ventaja computacional** sobre las variantes con FPN, tanto en entrenamiento como en inferencia.

Desde una perspectiva global, clínica y computacional, la variante CSWin-Transformer + ASPP en su versión base puede considerarse la configuración óptima del presente estudio.

6.8. Análisis cualitativo mediante técnicas visuales.

Además de las métricas cuantitativas presentadas en las secciones anteriores, se realizó un análisis cualitativo mediante técnicas visuales con el objetivo de observar qué regiones de la imagen influyeron en la respuesta de los modelos. Este análisis no debe interpretarse como una segmentación clínica de lesiones ni como una validación diagnóstica independiente, sino como una herramienta complementaria para revisar la coherencia espacial de las predicciones.

En este trabajo se utilizaron dos tipos de visualización: Grad-CAM y mapas de activación profunda. Grad-CAM permite resaltar regiones asociadas con la predicción de una clase específica, ya que utiliza los gradientes de la salida del modelo respecto a mapas internos de características. Por otro lado, los mapas de activación profunda permiten observar la intensidad de respuesta de una etapa interna del backbone, en este caso la etapa C4, sin depender directamente de los gradientes de una clase particular.

Aunque el CSWin-Transformer se fundamenta en mecanismos de atención, en esta sección se decidió reportar mapas de activación y Grad-CAM en lugar de mapas de atención crudos. Esta decisión se debe a que, en una arquitectura como CSWin, la atención se distribuye en múltiples cabezas, etapas y ventanas cruzadas horizontales y verticales. Además, la decisión final no depende únicamente del backbone, sino también del módulo multiescala empleado, ya sea FPN o ASPP, y de la cabeza de clasificación. Por ello, un mapa de atención aislado no necesariamente representa de forma directa la región que determinó la clase final. En cambio, Grad-CAM ofrece una aproximación más adecuada para analizar la contribución espacial asociada con la predicción, mientras que los mapas de activación profunda permiten revisar la respuesta interna del modelo ante las estructuras retinianas.

6.8.1. Grad-CAM

La Figura 6.13 muestra los mapas Grad-CAM obtenidos con la variante CSWin-Transformer + FPN. En los ejemplos correspondientes a CNV, DRUSEN y NORMAL, el modelo concentra su respuesta principalmente sobre la región retiniana visible. Esto sugiere que la clasificación se apoya en información anatómica relevante y no únicamente en el fondo de la imagen o en artefactos del preprocesamiento.

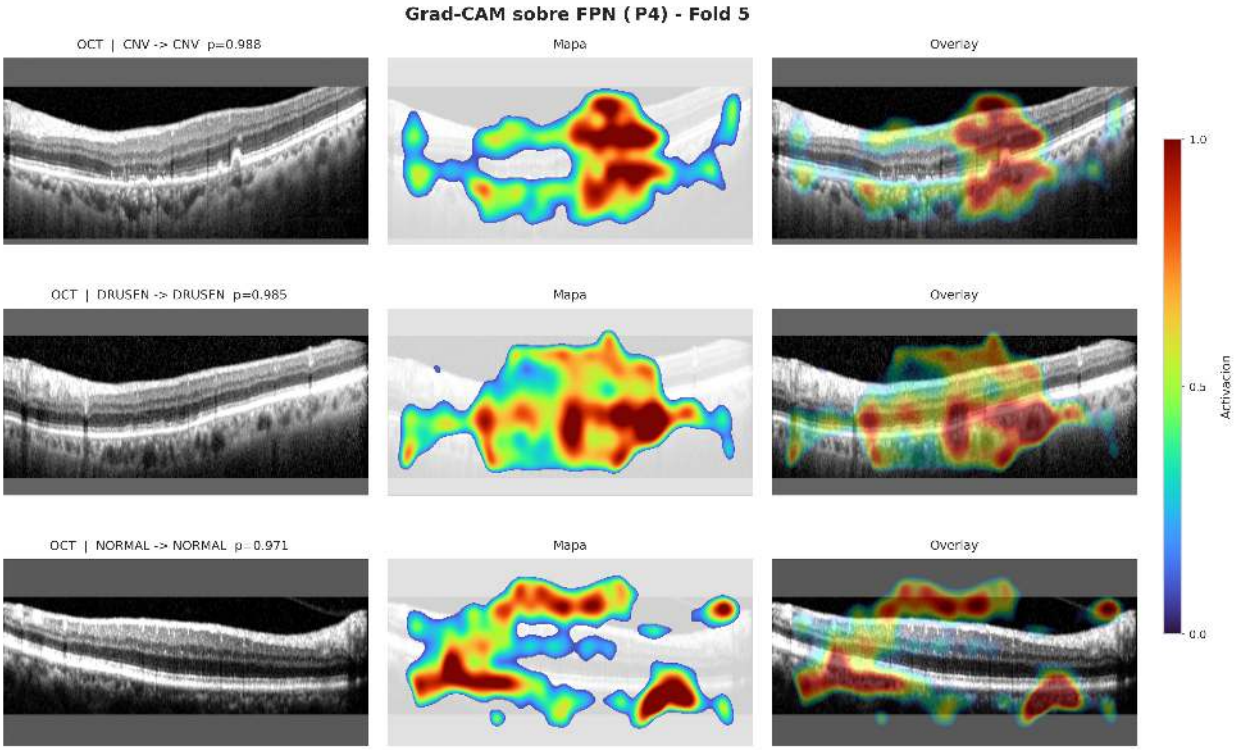


Figura 6.13: Representación visual del Grad-CAM, CSWin-Transformer + FPN.

En la clase CNV, las activaciones aparecen sobre zonas amplias e irregulares de la retina, lo cual es coherente con la presencia de alteraciones estructurales más extensas. En DRUSEN, la respuesta se distribuye alrededor de regiones asociadas con cambios en la línea del epitelio pigmentario y en la zona externa de la retina. Para la clase NORMAL, el mapa tiende a seguir regiones donde se conserva la continuidad de las capas retinianas, lo cual indica que el modelo también aprende patrones asociados con una arquitectura retinal preservada.

Sin embargo, en la variante con FPN se observa una distribución relativamente fragmentada de la activación. Algunas zonas de alta respuesta aparecen en regiones periféricas o se extienden de forma menos compacta sobre el B-scan. Esto puede relacionarse con la propia naturaleza de FPN, que fusiona información de distintos niveles jerárquicos y puede resaltar tanto detalles locales como regiones semánticas más amplias. Aunque esta estrategia permite recuperar información multiescala, también puede generar mapas visualmente más dispersos cuando se analizan mediante Grad-CAM.

La Figura 6.14 presenta los mapas Grad-CAM obtenidos con la variante CSWin-Transformer + ASPP. En comparación con FPN, las activaciones tienden a mostrarse más concentradas sobre regiones anatómicas continuas de la retina. Esto es especialmente evidente en CNV, donde el mapa resalta una zona amplia del tejido retiniano, y en DRUSEN, donde se observan focos de activación sobre regiones compatibles con cambios estructurales localizados.

Este comportamiento es coherente con el diseño del ASPP, ya que sus convoluciones dilatadas permiten integrar contexto a diferentes escalas sobre una representación profunda del backbone. En consecuencia, el modelo puede responder simultáneamente a patrones locales y a configuraciones

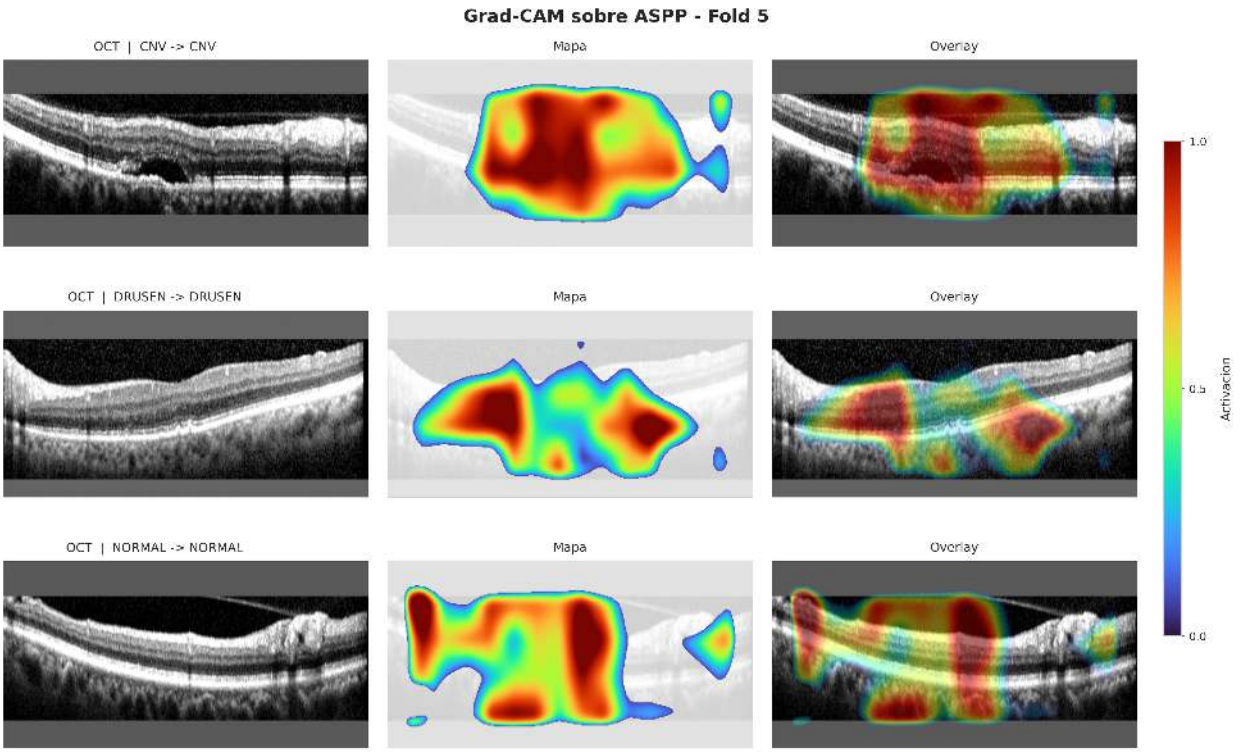


Figura 6.14: Representación visual del Grad-CAM, CSWin-Transformer + ASPP.

anatómicas más amplias. Esta integración contextual puede explicar por qué la variante con ASPP obtuvo mejores resultados globales que FPN en las métricas cuantitativas, particularmente en Macro-F1 y Macro-AUC.

En conjunto, los mapas Grad-CAM muestran que ambas arquitecturas toman decisiones a partir de regiones anatómicamente plausibles. No obstante, la variante con ASPP presenta una activación visualmente más compacta y mejor alineada con regiones retinianas relevantes, lo cual respalda cualitativamente su mejor comportamiento observado en la evaluación cuantitativa.

6.8.2. Mapas de activación profunda C4

La Figura 6.15 muestra los mapas de activación profunda de la etapa C4 para la variante CSWin-Transformer + FPN. A diferencia de Grad-CAM, estos mapas no representan directamente la contribución de una clase específica, sino la intensidad de respuesta de características profundas aprendidas por el backbone. Por ello, su interpretación debe centrarse en observar si la red activa regiones internas coherentes con el contenido anatómico de la OCT.

En los ejemplos presentados, las activaciones se ubican principalmente sobre la zona de la retina y siguen parcialmente su morfología. Esto indica que la etapa profunda C4 conserva información estructural relevante, aun cuando la resolución espacial ya se ha reducido por el procesamiento jerárquico del Transformer. En CNV y DRUSEN se observan zonas de activación asociadas con regiones donde existen variaciones morfológicas, mientras que en NORMAL la respuesta tiende a distribuirse sobre la continuidad de las capas retinianas.

No obstante, la activación en FPN muestra algunos focos más dispersos, lo cual puede deberse

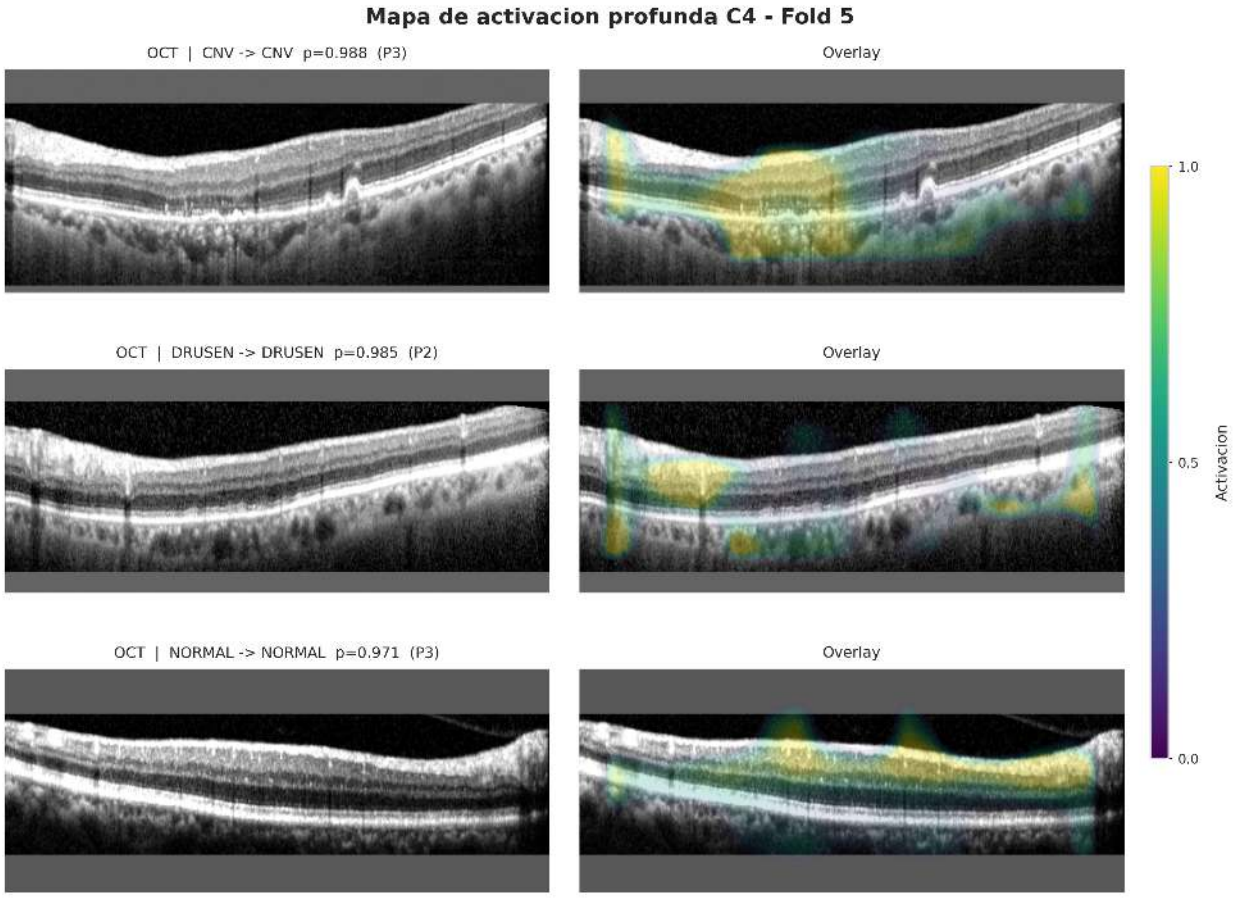


Figura 6.15: Representación visual del mapa de activación profunda C4, CSWin-Transformer + FPN.

a que esta variante combina distintos niveles de representación mediante el cuello piramidal. Esta característica favorece la integración multiescala, pero también puede producir respuestas espaciales menos compactas al visualizar una sola etapa profunda.

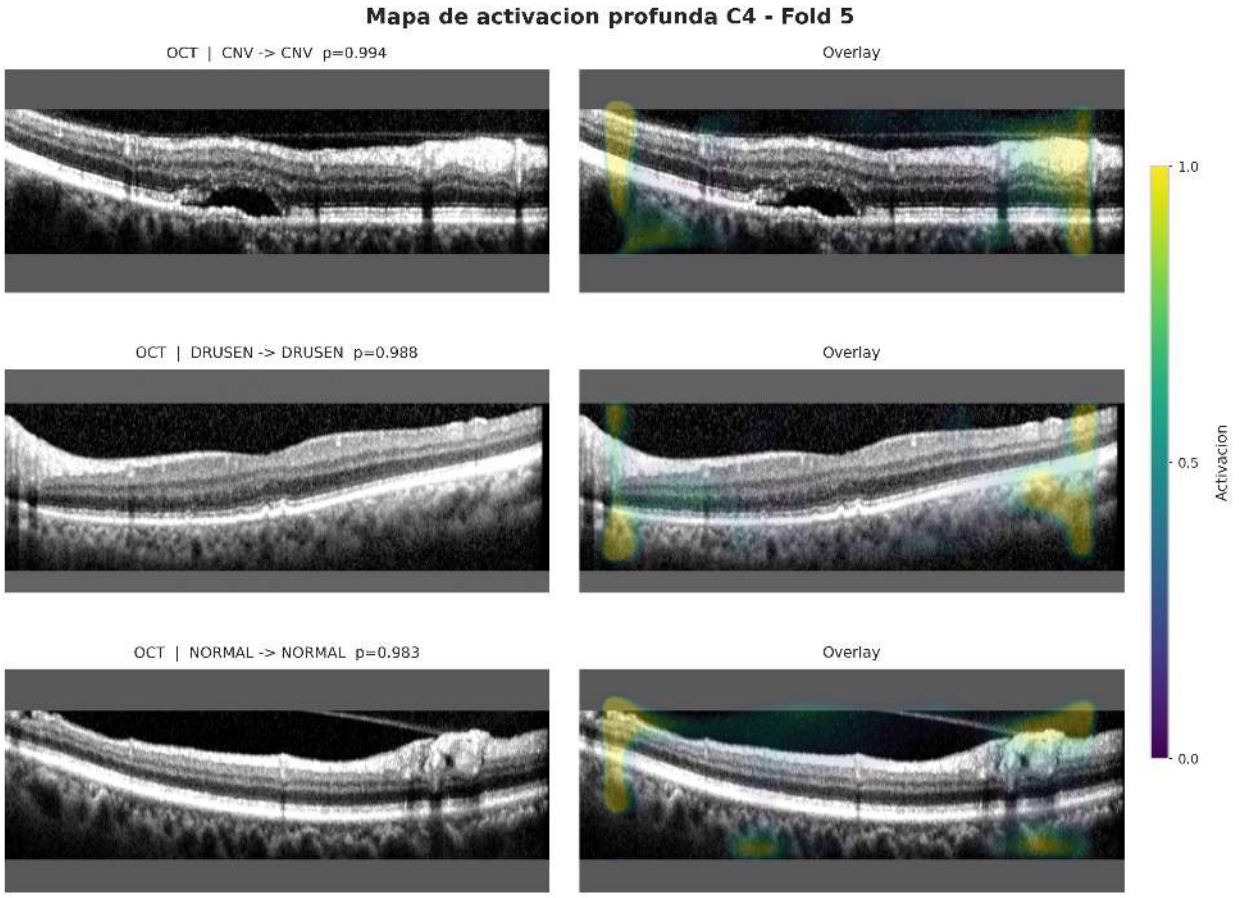


Figura 6.16: Representación visual del mapa de activación profunda C4, CSWin-Transformer + ASPP.

La Figura 6.16 muestra los mapas de activación profunda C4 para la variante CSWin-Transformer + ASPP. En este caso, las activaciones tienden a cubrir regiones continuas de la retina, manteniendo una respuesta más homogénea sobre zonas anatómicas relevantes. Esta distribución sugiere que la representación profunda utilizada por ASPP conserva información contextual amplia antes de la clasificación final.

En la clase CNV, la activación se extiende sobre una región amplia del B-scan, lo cual es consistente con la naturaleza más extensa e irregular de las alteraciones asociadas a esta categoría. En DRUSEN, la respuesta se concentra en zonas más focales, compatibles con cambios localizados en la arquitectura externa de la retina. En NORMAL, la activación se distribuye sobre áreas donde se observa continuidad de las capas retinianas, lo que sugiere que el modelo también utiliza la preservación estructural como evidencia para clasificar imágenes sin hallazgos patológicos relevantes.

En términos generales, los mapas de activación profunda complementan la interpretación obtenida mediante Grad-CAM. Mientras Grad-CAM permite observar regiones asociadas con la clase

predicha, los mapas C4 muestran que el backbone genera respuestas internas sobre zonas anatómicamente coherentes. La combinación de ambas visualizaciones refuerza la interpretación de que los modelos no se apoyan únicamente en artefactos externos o regiones de fondo, sino en patrones estructurales presentes en la retina.

Finalmente, el análisis cualitativo es consistente con los resultados cuantitativos del estudio. La variante CSWin-Transformer + ASPP no sólo alcanzó el mejor desempeño global, sino que también mostró mapas visuales más concentrados y coherentes con regiones retinianas relevantes. Por esta razón, las visualizaciones apoyan la selección de ASPP como la configuración final más adecuada dentro de los experimentos realizados.

6.9. Modelo Final : CSWin-Transformer + ASPP.

6.9.1. Selección del modelo óptimo.

Tras la evaluación comparativa entre las arquitecturas propuestas, tanto en sus versiones base como después del proceso de Fine-Tuning, se determinó que el modelo **CSWin-Transformer + ASPP (versión base)** constituye la configuración con mejor desempeño global y mayor estabilidad discriminativa.

La selección se fundamenta en los siguientes criterios:

- Mayor **exactitud global**: 0.9440 ± 0.0073.
- Mejor **Macro-F1**: 0.9433 ± 0.0069, lo que garantiza un desempeño balanceado entre clases.
- Mayor **Weighted-F1**: 0.9436 ± 0.0072, consistente con una respuesta robusta a nivel global.
- Superior **Macro-AUC**: 0.9799 ± 0.0014, evidenciando mayor capacidad de separación probabilística entre clases.
- **Micro-AUC** máximo de 0.9811, compartido con la variante ASPP + FT, pero con menor desviación estándar en la versión base.
- Curvas ROC y matriz de confusión coherentes con un patrón de menor confusión cruzada y mejor equilibrio general entre categorías.

En conjunto, estos resultados indican que la arquitectura no sólo clasifica correctamente una mayor proporción de imágenes, sino que también asigna probabilidades mejor separadas y mantiene un comportamiento más estable entre folds.

6.9.2. Fundamentos de desempeño superior basados en la arquitectura.

El rendimiento superior del modelo CSWin-Transformer + ASPP puede explicarse desde una perspectiva estructural y funcional.

1. Backbone CSWin-Transformer.

El CSWin-Transformer introduce atención cruzada en ventanas (*Cross-Shaped Window Attention*), lo cual permite:

- Capturar dependencias espaciales de largo alcance.
- Mantener eficiencia computacional al restringir la atención a particiones estructuradas.
- Preservar la jerarquía multiescala mediante etapas progresivas.

En imágenes OCT, donde las alteraciones patológicas pueden extenderse horizontalmente a lo largo de capas retinianas completas, este mecanismo resulta particularmente adecuado para modelar patrones estructurales complejos y relaciones espaciales amplias.

2. Integración multiescala mediante ASPP.

El módulo *Atrous Spatial Pyramid Pooling* (ASPP) incorpora convoluciones dilatadas con múltiples tasas de dilatación, lo que permite:

- Expandir el campo receptivo efectivo sin incrementar de forma importante el número de parámetros.
- Capturar patrones locales y globales simultáneamente.
- Mejorar la sensibilidad ante variaciones morfológicas de distinto tamaño.

En el contexto de la DMAE, donde las lesiones pueden presentarse como drusas pequeñas, alteraciones intermedias o neovascularizaciones más extensas, esta capacidad multiescala resulta especialmente valiosa para capturar tanto micro-patrones locales como contexto anatómico de mayor amplitud.

6.9.3. Comparación con FPN.

Aunque la *Feature Pyramid Network* (FPN) también integra información multiescala, su mecanismo se basa en la combinación lateral de mapas jerárquicos mediante una ruta *top-down*. En contraste, ASPP realiza una exploración paralela del espacio de características por medio de dilataciones controladas sobre una representación de alto nivel.

Mientras FPN prioriza la fusión jerárquica entre niveles del backbone, ASPP enfatiza la exploración contextual multirresolución dentro de una misma representación semánticamente rica. En este problema específico de clasificación OCT, la captación contextual amplia proporcionada por ASPP demostró mayor efectividad que la agregación piramidal de FPN, tanto en términos de exactitud como de capacidad discriminativa y eficiencia de inferencia.

6.9.4. Análisis de estabilidad y generalización.

Un aspecto clave en la selección del modelo final fue la estabilidad observada entre folds. La variante CSWin-Transformer + ASPP base presentó:

- Curvas de entrenamiento con convergencia estable.
- Baja variabilidad en las métricas globales, especialmente en *Macro-AUC* y *Micro-AUC*.
- Buen equilibrio entre desempeño global y desempeño interclase.
- Inferencia eficiente en GPU A100, con 144.2607 ± 0.8209 imágenes por segundo.

Además, el hecho de que el Fine-Tuning no aportara mejoras claras sobre ASPP base refuerza la idea de que el modelo ya había alcanzado una representación altamente adecuada desde la fase inicial de entrenamiento. En otras palabras, la variante base no sólo fue la mejor en términos absolutos, sino también la más robusta frente a refinamientos posteriores.

Por lo tanto, el modelo CSWin-Transformer + ASPP se establece como el modelo final del presente trabajo, al ofrecer:

1. El mejor equilibrio entre exactitud y robustez interclase.
2. Mayor capacidad de modelado contextual multiescala.
3. Mejor discriminación probabilística entre categorías.
4. Estabilidad durante entrenamiento, validación y evaluación.
5. Alta eficiencia computacional en GPU A100.

Este modelo sintetiza de manera óptima la atención global jerárquica del CSWin-Transformer con la exploración contextual multiescala del ASPP, resultando particularmente adecuado para la clasificación automatizada de imágenes OCT orientada al diagnóstico de DMAE.

6.10. Comparación con el estado del arte.

Tabla 6.21: Comparación integral frente al estado del arte en clasificación OCT (NORMAL/DRUSEN/CNV). En el caso de esta tesis se reportan los resultados promedio y la desviación estándar obtenidos mediante validación cruzada estratificada y agrupada de 5 folds.

Trabajo	Arquitectura	Multiescala	Protocolo	Accuracy	Macro-F1	Macro-AUC
Kermany, Goldbaum et al., 2018	Inception-V3 (CNN)	No	Train/Test predefinido	96.53 %	–	–
Hassan et al., 2023	ResNet-50 + RF	No	Train/Test	97.56 %	0.9688	–
Sotoudeh-Paima et al., 2022	VGG16 + FPN	Sí (FPN)	CV 5-fold (paciente)	92.0 %	–	–
Yusufoğlu et al., 2024	EfficientNetB0 + Atención (MSA-Net)	Sí	90/10 sin val	98.1 %	0.98 (global)	–
Tesis	**CSWin + ASPP**	**Sí (ASPP)**	**CV 5-fold estratificada y agrupada + test externo fijo**	94.40 ± 0.73 **%**	0.9433 ± 0.0069	0.9799 ± 0.0014

La Tabla 6.21 sitúa los resultados de la presente tesis frente a trabajos representativos en clasificación de imágenes OCT. En este caso, se tomó como referencia principal el desempeño promedio obtenido mediante validación cruzada estratificada y agrupada de 5 folds, acompañado

de su desviación estándar, ya que este criterio permite valorar no sólo el nivel de rendimiento alcanzado por el modelo, sino también su estabilidad frente a distintas particiones del conjunto de desarrollo. A diferencia de una única partición fija, este enfoque ofrece una estimación más robusta del comportamiento esperado del sistema.

Dentro del análisis interno de esta tesis, la variante **CSWin-Transformer + ASPP** fue la que mostró el mejor comportamiento global entre las arquitecturas evaluadas, superando de forma consistente a **CSWin-Transformer + FPN**. En particular, el promedio alcanzado por **CSWin + ASPP** fue de $94.40 \pm 0.73\,\%$ en exactitud, con un *Macro-F1* de 0.9433 ± 0.0069 y un *Macro-AUC* de 0.9799 ± 0.0014. En conjunto, estos resultados sugieren que la incorporación de ASPP favorece una integración contextual multiescala más efectiva dentro de la arquitectura híbrida propuesta.

Al contrastar estos hallazgos con trabajos previos, se observa que la propuesta desarrollada en esta tesis se mantiene en un rango competitivo frente a métodos relevantes de la literatura. Por ejemplo, supera con claridad la exactitud reportada por Sotoudeh-Paima et al., 2022, quienes evaluaron una configuración basada en VGG16+FPN sobre la misma problemática y alcanzaron aproximadamente $92.0\,\%$ de exactitud. Esta diferencia sugiere que la combinación de un backbone Transformer jerárquico con un módulo multiescala especializado puede ofrecer ventajas frente a enfoques convolucionales piramidales más convencionales.

La comparación con Kermany, Goldbaum et al., 2018 y Hassan et al., 2023 debe interpretarse con cautela, ya que ambos estudios utilizaron configuraciones experimentales distintas. En el caso de Kermany, Goldbaum et al., 2018, la tarea abordada consideró cuatro clases, por lo que no resulta estrictamente equivalente al problema tratado en esta tesis. En cuanto a Hassan et al., 2023, aunque reporta un desempeño elevado, su evaluación se realizó bajo un esquema de partición simple y sin el mismo nivel de control experimental adoptado en el presente trabajo.

Un caso particularmente relevante es Yusufoğlu et al., 2024, cuyo modelo MSA-Net reportó una exactitud de $98.1\,\%$. Aunque este valor es superior al promedio obtenido en esta tesis, la comparación no es completamente directa, ya que dicho estudio utilizó una partición 90/10 sin validación cruzada agrupada por paciente ni un conjunto de prueba externo fijo independiente. En contraste, el presente trabajo adoptó una estrategia metodológica más estricta, basada en separación por paciente, prueba externa fija y validación cruzada estratificada y agrupada de 5 folds, lo que reduce el riesgo de fuga de información y proporciona una estimación más robusta del desempeño.

En este sentido, la principal aportación del presente trabajo no radica únicamente en alcanzar resultados competitivos, sino en demostrar que una arquitectura híbrida basada en **CSWin-Transformer + ASPP** puede sostener un desempeño sólido bajo un protocolo de evaluación más exigente y metodológicamente más controlado. Desde esta perspectiva, la comparación con el estado del arte sugiere que la propuesta no sólo es eficaz en términos de clasificación, sino también consistente desde el punto de vista experimental, al mantener su rendimiento bajo condiciones de validación más rigurosas.

Cabe señalar que la comparación directa del costo computacional con trabajos previos es limitada, ya que no se reportan datos temporales sobre el hardware utilizado, la latencia por imagen o el tiempo de entrenamiento. Por esta razón, en el presente trabajo el análisis computacional se centró en una comparación controlada entre las variantes propuestas bajo el mismo entorno experimental.

CAPÍTULO 7

Conclusiones.

El presente trabajo tuvo como objetivo desarrollar y evaluar un sistema de aprendizaje profundo para la clasificación automática de imágenes de Tomografía de Coherencia Óptica (OCT) en tres categorías clínicas asociadas con la Degeneración Macular Asociada a la Edad (DMAE): **NORMAL**, **DRUSEN** y **CNV**. Para ello, se diseñó, implementó y analizó una propuesta basada en un *backbone* jerárquico **CSWin-Transformer**, sobre el cual se compararon dos estrategias de integración multiescala: **Feature Pyramid Network (FPN)** y **Atrous Spatial Pyramid Pooling (ASPP)**.

En relación con el diseño de la arquitectura, los resultados mostraron que la combinación entre **CSWin-Transformer** y un módulo multiescala fue adecuada para modelar patrones retinianos complejos en OCT. La estructura jerárquica del backbone permitió conservar información en distintos niveles de representación, mientras que la comparación entre FPN y ASPP hizo posible analizar de forma clara el efecto del módulo de agregación sobre el desempeño final. Dentro de este marco, la variante **CSWin-Transformer + ASPP** fue la que mostró el comportamiento más sólido y consistente.

Respecto a la obtención y organización de la base de datos, una parte importante del trabajo no consistió únicamente en utilizar el conjunto disponible, sino en reorganizarlo bajo un protocolo experimental más estricto, centrado en la separación por paciente. Esta decisión permitió reducir el riesgo de fuga de información entre subconjuntos y sentó una base metodológica más confiable para la evaluación de los modelos.

En cuanto al preprocesamiento, se construyó una tubería orientada específicamente a OCT monocanal, integrada por recorte guiado por región de interés, estandarización geométrica mediante *padding* reflectivo y *letterbox*, así como normalización *z-score* dentro de la región válida. Este esquema permitió homogeneizar la entrada del modelo, reducir variaciones no clínicas y conservar la anatomía útil de cada B-scan. De manera complementaria, la aplicación controlada de aumentación geométrica y fotométrica sólo sobre el conjunto de entrenamiento ayudó a mejorar la generalización del modelo sin comprometer la coherencia anatómica de las imágenes.

En lo referente a la evaluación de la eficacia del método propuesto, los resultados confirmaron la hipótesis de trabajo. La variante **CSWin-Transformer + ASPP**, en su configuración base,

alcanzó el mejor desempeño global del estudio, con una exactitud de 0.9440 ± 0.0073, un *Macro-F1* de 0.9433 ± 0.0069, un *Weighted-F1* de 0.9436 ± 0.0072, un *Macro-AUC* de 0.9799 ± 0.0014 y un *Micro-AUC* de 0.9811 ± 0.0011. Estos resultados muestran no sólo una alta proporción de clasificaciones correctas, sino también un comportamiento equilibrado entre clases y una buena capacidad de separación probabilística.

La comparación entre arquitecturas mostró de forma consistente que **ASPP** superó a **FPN** dentro del mismo backbone y bajo el mismo protocolo experimental. Esto sugiere que, para el problema abordado, la agregación contextual mediante convoluciones dilatadas en paralelo resultó más efectiva que la fusión piramidal tradicional para capturar al mismo tiempo patrones locales y contexto anatómico de mayor alcance. En otras palabras, el principal hallazgo técnico de esta tesis no fue sólo que el modelo funcionara bien, sino que la variante con **ASPP** ofreció el mejor equilibrio entre desempeño, estabilidad y capacidad de generalización.

En cuanto al ajuste y optimización del modelo, el análisis del *fine-tuning* mostró un comportamiento distinto entre arquitecturas. En la variante con FPN, el ajuste fino produjo una mejora apreciable, lo que sugiere que aún existía margen para refinar la representación aprendida en la fase base. En cambio, en la variante con ASPP las ganancias fueron menores, lo que indica que el modelo ya alcanzaba desde la etapa base una representación suficientemente discriminativa. Este resultado refuerza la idea de que **CSWin-Transformer + ASPP** constituye la configuración más estable y efectiva del estudio.

Desde el punto de vista metodológico, una de las aportaciones más importantes del trabajo fue la adopción de un esquema experimental más riguroso que el utilizado en varios antecedentes. La combinación de un conjunto de prueba externo fijo con validación cruzada estratificada y agrupada de 5 folds permitió reducir la dependencia de una sola partición, controlar de manera explícita la separación por paciente y obtener estimaciones promedio del desempeño más robustas. En este sentido, la aportación de la tesis no se limita al valor numérico alcanzado, sino también a la forma en que dicho desempeño fue demostrado.

En términos de interpretabilidad, la incorporación de técnicas de análisis visual como Grad-CAM y mapas de atención aporta un elemento adicional de valor al sistema propuesto, ya que permite examinar si las predicciones del modelo se apoyan en regiones anatómicamente plausibles. Aunque estas herramientas no sustituyen la evaluación cuantitativa, sí fortalecen la lectura cualitativa del comportamiento del modelo y ayudan a entender mejor cómo toma decisiones la red.

De manera global, puede concluirse que la arquitectura **CSWin-Transformer + ASPP**, en su versión base, representa la mejor configuración obtenida en esta investigación. Su desempeño cuantitativo, su estabilidad entre folds y su comportamiento competitivo frente al estado del arte indican que constituye una alternativa sólida para la clasificación automática de imágenes OCT relacionadas con DMAE.

No obstante, el trabajo también presenta limitaciones. La evaluación se realizó sobre una sola base de datos, no se incorporó validación externa multicéntrica y el enfoque se centró en clasificación global, sin segmentación explícita de biomarcadores retinianos. Además, aunque el modelo mostró resultados favorables, su posible despliegue en dispositivos con recursos limitados o entornos *edge* requeriría etapas adicionales de optimización, como compresión del modelo, cuantización

o reducción de complejidad computacional.

Como líneas futuras, resulta pertinente extender la validación a bases independientes y multicéntricas, incorporar estrategias de calibración probabilística, explorar esquemas multitarea que combinen clasificación y segmentación, y profundizar en mecanismos de explicabilidad que faciliten la interpretación clínica de las decisiones del modelo. Del mismo modo, sería valioso estudiar variantes más ligeras o estrategias de aceleración que acerquen este tipo de sistemas a aplicaciones de apoyo diagnóstico en contextos de infraestructura limitada.

En síntesis, esta tesis aporta evidencia de que una arquitectura híbrida basada en **CSWin-Transformer + ASPP**, acompañada de un preprocesamiento orientado a OCT y de un protocolo de evaluación estricto a nivel de paciente, puede ofrecer una solución eficaz, robusta y metodológicamente sólida para la clasificación automática de DMAE en imágenes OCT.

Referencias

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2022). *Molecular Biology of the Cell* (7.ª ed.). W. W. Norton & Company.

Arsalan, M., et al. (2022). Automated Deep Learning-based Segmentation of OCT Images for Retinal Disease Diagnosis. *Medical Image Analysis*, *82*, 102308. https://doi.org/10.1016/j.media.2022.102308

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems (NeurIPS)*, *33*, 1877-1901.

Cabaleiro, P., de Moura, J., Novo, J., Charlón, P., & Ortega, M. (2019). Automatic Identification and Representation of the Cornea–Contact Lens Relationship Using AS-OCT Images. *Sensors*, *19*(23), 5087. https://doi.org/10.3390/s19235087

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 834-848. https://doi.org/10.1109/TPAMI.2017.2699184

Chen, L.-C., Papandreou, G., Schroff, F., & Adam, H. (2018). Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv preprint arXiv:1706.05587*. http://arxiv.org/abs/1706.05587

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *European Conference on Computer Vision (ECCV)*, 833-851. https://doi.org/10.1007/978-3-030-01234-2_49

Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. *CVPR*, 1251-1258.

De la Rosa, F. (2024). Aplicación de deep learning en robótica móvil para exploración y reconocimiento de objetos basados en imágenes [Accedido el 15 de mayo de 2024]. Consultado el 15 de mayo de 2024, desde https://www.researchgate.net/figure/Red-neuronal-convolucional-4_fig7_308783857

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*. http://arxiv.org/abs/1810.04805

Dong, X., Bao, J., Chen, D., Zhang, W., Chen, L., Yuan, L., & Wen, F. (2022). CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. *arXiv preprint arXiv:2107.00652*.

Drexler, W., & Fujimoto, J. G. (2008). Optical Coherence Tomography–Principles and Applications. *Reports on Progress in Physics*, *71*(4), 046601.

Franco, E., & Ramos, R. (2019). Aprendizaje de máquina y aprendizaje profundo en biotecnología: aplicaciones, impactos y desafíos. *Ciencia, Ambiente y Clima*, *2*, 7-26. https://doi.org/10.22206/cac.2019.v2i2.pp7-26

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org

Hassan, E., Elmougy, S., Ibraheem, M. R., Hossain, M. S., AlMutib, K., Ghoneim, A., AlQahtani, S. A., & Talaat, F. M. (2023). Enhanced Deep Learning Model for Classification of Retinal Optical Coherence Tomography Images. *Sensors*, *23*(12), 5393. https://doi.org/10.3390/s23125393

He, K., Zhang, X., Ren, S., & Sun, J. (2016a). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778.

He, K., Zhang, X., Ren, S., & Sun, J. (2016b). Deep Residual Learning for Image Recognition. *CVPR*, 770-778.

Howard, A., Sandler, M., Chu, G., et al. (2019). Searching for MobileNetV3. *ICCV*, 1314-1324.

Howard, A. G., Zhu, M., Chen, B., et al. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv:1704.04861*.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-Excitation Networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7132-7141. https://doi.org/10.1109/CVPR.2018.00745

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. *CVPR*, 4700-4708.

Instituto Diagonal. (n.d.). Hipermetropía [Recuperado el 15 de octubre de 2025].

Kadir, T., et al. (2023). EdgeAL: An Edge Estimation Based Active Learning Approach for OCT Segmentation. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-023-10345-6

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., & Baxter. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, *172*(5), 1122-1131. https://doi.org/10.1016/j.cell.2018.02.010

Kermany, D. S., Zhang, K., & Goldbaum, M. (2018). Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification [Dataset]. https://doi.org/10.17632/rscbjbr9sj.1

Khurana, A. K. (2023). *Comprehensive Ophthalmology* (7.ª ed.). Jaypee Brothers Medical Publishers.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, *25*, 1097-1105.

Ledesma-Carbayo, M. J., et al. (2023). Robust Deep Learning-based Approach for Retinal Layer Segmentation in OCT. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-022-10345-6

Lim, J. I. (Ed.). (2013). *Age-Related Macular Degeneration* (3.ª ed.). CRC Press.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature Pyramid Networks for Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2117-2125. https://doi.org/10.1109/CVPR.2017.106

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J., van Ginneken, B., & Sánchez, C. I. (2017). A Survey on Deep Learning in Medical Image Analysis. *Medical Image Analysis*, *42*, 60-88.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. https://arxiv.org/abs/2103.14030

MedlinePlus. (2022, abril). *Degeneración macular asociada con la edad* [Biblioteca Nacional de Medicina de EE. UU.]. https://medlineplus.gov/spanish/ency/article/001000.htm

National Eye Institute. (2022). Age-related Macular Degeneration (AMD). https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/age-related-macular-degeneration

Oftalvist. (n.d.). Degeneración macular [Recuperado el 15 de octubre de 2025].

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 234-241.

Ruiz Casas, D. (n.d.). Retina – Anatomía del globo ocular [Recuperado el 15 de octubre de 2025].

Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4.ª ed.). Pearson.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. *CVPR*, 4510-4520.

Scanner Vizcaya. (2017). Tomografía de coherencia óptica [Imagen disponible en línea]. Consultado el 15 de mayo de 2024, desde https://www.gruposcanner.biz/otc-bilbao/

Shi, Y., et al. (2023). Advanced OCT Image Segmentation Using Multi-Scale Deep Learning. *Medical Image Analysis*, *83*, 104567. https://doi.org/10.1016/j.media.2023.104567

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *ICLR*.

Sotoudeh-Paima, S., Hajizadeh, F., & Soltanian-Zadeh, H. (2023). Labeled Retinal Optical Coherence Tomography Dataset for Classification of Normal, Drusen, and CNV Cases [Data set]. https://doi.org/10.17632/8kt969dhx6.2

Sotoudeh-Paima, S., Jodeiri, A., Hajizadeh, F., & Soltanian-Zadeh, H. (2022). Multi-Scale Convolutional Neural Network for Automated AMD Classification Using Retinal OCT Images. *Computers in Biology and Medicine*, *144*, 105368. https://doi.org/10.1016/j.compbiomed.2022.105368

Szegedy, C., Liu, W., Jia, Y., et al. (2015). Going Deeper with Convolutions. *CVPR*, 1-9.

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *ICML*, 6105-6114.

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: A Systematic Review and Meta-Analysis. *Ophthalmology*, *121*(11), 2081-2090. https://doi.org/10.1016/j.ophtha.2014.05.013

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems (NeurIPS)*, 5998-6008.

Wong, W. L., Su, X., Li, X., Cheung, C. M. G., Klein, R., Cheng, C. Y., & Wong, T. Y. (2014). Global Prevalence of Age-Related Macular Degeneration and Disease Burden Projection for

2020 and 2040: A Systematic Review and Meta-Analysis. *The Lancet Global Health*, *2*(2), e106-e116. https://doi.org/10.1016/S2214-109X(13)70145-1

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional Block Attention Module. *European Conference on Computer Vision (ECCV)*, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

World Health Organization. (2019). *World Report on Vision*. https://www.who.int/publications/i/item/world-report-on-vision

Yusufoğlu, E., Fırat, H., Üzen, H., Özçelik, S. T. A., Balıkçı Çiçek, İ., Şengür, A., Atila, O., & Guldemir, N. H. (2024). A Comprehensive CNN Model for Age-Related Macular Degeneration Classification Using OCT: Integrating Inception Modules, SE Blocks, and ConvMixer. *Diagnostics*, *14*(24), 2836. https://doi.org/10.3390/diagnostics14242836