ESTADIFICACIÓN IMAGENOLÓGICA DE LA ATROFIA GEOGRÁFICA EN LA DEGENERACIÓN MACULAR RELACIONADA CON LA EDAD (DMAE), UTILIZANDO TOMOGRAFÍA ÓPTICA COHERENTE (OCT) CON 3 CORTES, EN UNA POBLACIÓN DEL SUR DE COLOMBIA

Andrés Felipe Quiñones Lucio
Código 5144203

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director
Ing. Hernán Darío Vargas Cardona, PhD

Codirectora
Dra. Diana Marcela Libreros Arango, Oftalmóloga - Retinóloga

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, DICIEMBRE DE 2024

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE GRÁFICOS

INTRODUCCIÓN

La degeneración macular atrófica relacionada con la edad (DMAE), es una enfermedad degenerativa que afecta el órgano de la visión en su segmento posterior. Puntualmente, se afecta la parte central de la retina que es responsable de la visión detallada y nítida, llamada mácula. Esta patología, es una de las principales causas de pérdida visual en personas mayores de 50 años [3]. Fisiopatológicamente, se produce cuando las células de la mácula inician un proceso de descomposición y muerte (apoptosis), que gradualmente redunda en detrimento de la visión central [1]. Esto ocurre por acumulación de depósitos de desechos celulares debajo de la retina (especialmente en la mácula) generando la formación de una zona de degeneración y produciendo disminución en la función de las células retinianas [1].

La estadificación de la atrofia geográfica en la degeneración macular asociada con la edad (DMAE) es fundamental para la comprensión de la progresión de la enfermedad la cual apunta a obtener mejor diagnóstico y pronóstico de la misma y guía para la toma de decisiones con relación al tratamiento o a estudios futuros que enfoquen posibles tratamientos para la DMAE. Es así como, la tomografía óptica coherente (OCT) del segmento posterior del ojo, proporciona imágenes detalladas, que dependiendo de su resolución aportan datos que se convierten en información a la hora de aplicar técnicas de ciencia de datos en su análisis [4].

Este trabajo de grado de maestría tuvo como norte desarrollar un modelo de aprendizaje de máquinas en el ámbito del aprendizaje profundo que permitiera estadificar la DMAE a través de las imágenes del OCT de un equipo de alta definición en tres cortes horizontales. Se realizó con las imágenes tomadas en una población del sur de Colombia con rango de edad entre los 50 y 99 años. Se tomó la decisión de utilizar el equipo OCT como imagen de alta tecnología y resolución, dado que, primero, está indicado en el estudio imagenológico de la patología y, segundo, permite la detección temprana de cambios en las capas externas de la estructura histológica de la retina. Esto, permitiría una evaluación precisa de la DMAE en la población objeto de estudio. Se ejecutó un análisis cuantitativo y cualitativo de las imágenes en busca de identificar patrones de atrofia geográfica y se estableció un sistema de estadificación.

Los resultados del presente trabajo pretenden impactar principalmente en la generación de información útil para la precisión diagnóstica y pronóstico de pacientes con esta patología, y apoyará futuros estudios enfocados en el tratamiento. Pretende adicionalmente, contribuir al avance del conocimiento científico en el campo de la ciencia de datos aplicada a la oftalmología.

1. DEFINICIÓN DEL PROBLEMA

La degeneración macular atrófica relacionada con la edad (DMAE) es una enfermedad que afecta la mácula, la región central de la retina donde se enfoca la luz. Esta zona es responsable de la visión detallada, medida por la agudeza visual (AV) durante la consulta de oftalmología u optometría, y su degeneración ocurre en personas mayores de 50 años. Actualmente, existen herramientas diagnósticas como la tomografía de coherencia óptica (OCT) que permiten visualizar el grado de atrofia macular; sin embargo, su interpretación requiere de un especialista en retina entrenado. Además, es necesaria la correlación clínica que el médico tratante (oftalmólogo) realiza entre el cuadro clínico del paciente, la imagen del OCT y la interpretación realizada por un retinólogo.

1.1. PLANTEAMIENTO DEL PROBLEMA

La DMAE es una enfermedad ocular crónica de presentación progresiva [13] que afecta principalmente a personas mayores de 50 años [1,2]. Esta condición se caracteriza por la degeneración de la mácula, parte de la retina responsable de la visión central y aguda. Uno de los subtipos más comunes de DMAE es la atrofia geográfica, que se caracteriza por la presencia de áreas de degeneración y adelgazamiento de la retina con presencia de sedimentación y acumulación de depósitos celulares y lipídicos llamados drusas [1,3], que son la piedra angular de la patología. A continuación, se observa la morfología normal del fondo del ojo humano (Figura No. 1), su correlación con una imagen de OCT macular (Figura No. 2), y la progresión de la atrofia geográfica en el tiempo y su impacto en la retina (Figura No. 3).

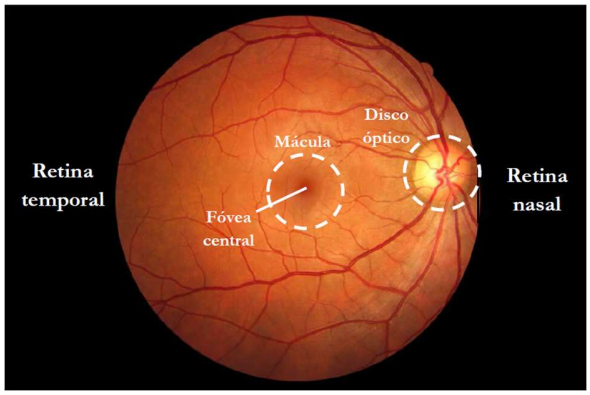


Figura No. 1 – Fondo de ojo humano derecho. [54].

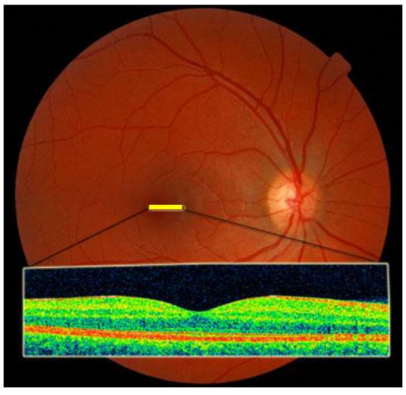


Figura No. 2 – Fondo de ojo y OCT macular. [54].

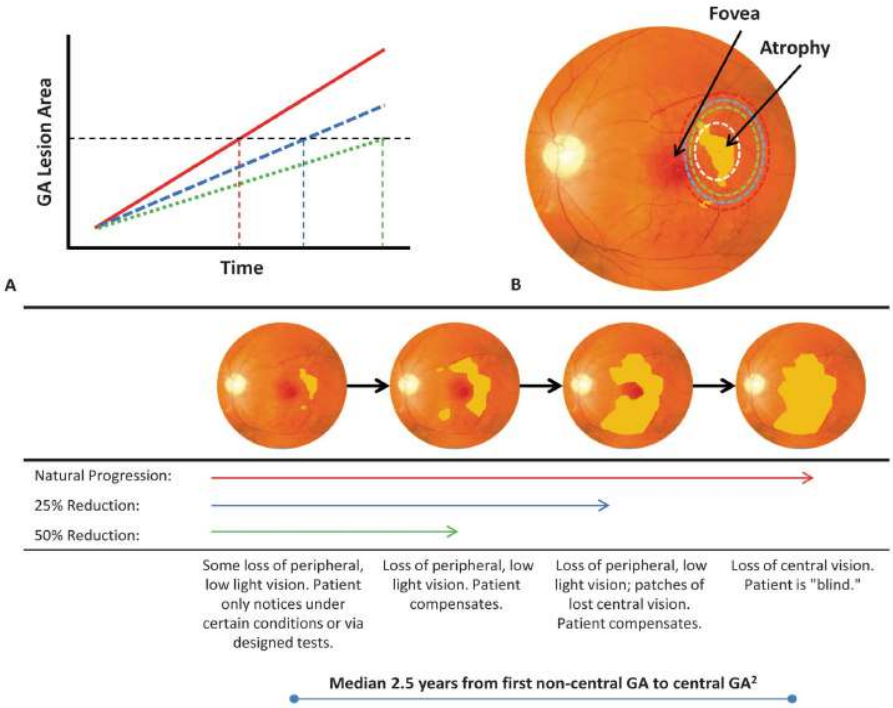


Figura No. 3 – Impacto visual de la progresión de la atrofia geográfica. *Dado que la progresión de la atrofia geográfica (AG) suele comenzar fuera de la fóvea, disminuir la tasa de progresión del área de AG entre un 25 % y un 50 % puede potencialmente retrasar la progresión hacia la fóvea durante años, especialmente si la intervención se inicia temprano.* ***Rojo****: Progresión natural;* ***azul****: reducción del 25 %;* ***verde****: reducción del 50 %.* [13].

Es fundamental estadificar la atrofia geográfica en la DMAE para así comprender la progresión de la enfermedad, evaluar su impacto en la calidad de vida de los pacientes [3,13] y desarrollar estrategias de tratamiento más efectivas. Actualmente, la tomografía de coherencia óptica (OCT) se ha convertido en una herramienta clave en la evaluación y diagnóstico de la DMAE, ya que permite obtener imágenes en alta resolución de la retina y sus capas [3,6] como se aprecia en la siguiente figura No. 4.

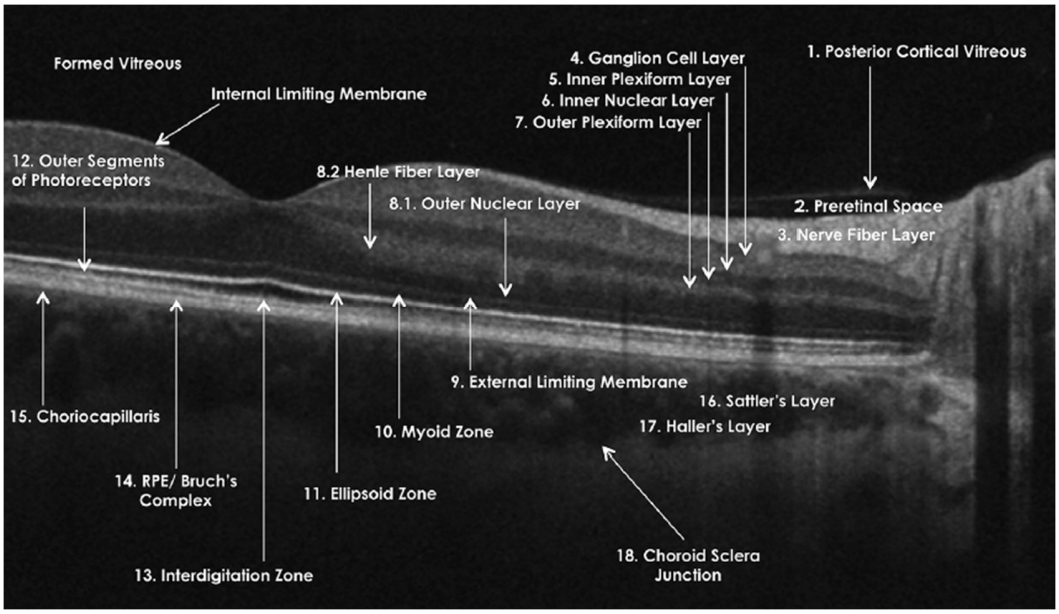


Figura No. 4 – Nomenclatura para los puntos de referencia anatómicos normales observados en imágenes de tomografía de coherencia óptica (OCT) de dominio espectral, propuesta y adoptada por el Panel Internacional de Nomenclatura para la Tomografía de Coherencia Óptica. Retina saludable capturada usando Zeiss Cirrus. RPE = epitelio pigmentario de la retina. [53]. *1. Capa cortical vítrea posterior, 2. Espacio prerretiniano, * Membrana limitante interna, 3. Capa de fibras nerviosas, 4. Capa de células ganglionares, 5. Capa plexiforme interna, 6. Capa nuclear interna, 7. Capa plexiforme externa, 8.1. Capa nuclear externa, 8.2. Capa de fibras de Henle, 9. Membrana limitante externa, 10. Zona mioidea, 11. Zona elipsoidea, 12. Segmento externo de fotorreceptores, 13. Zona de interdigitación, 14. RPE/Complejo de Bruch, 15. Capa coriocapilar, 16. Capa de Sattler, 17. Capa de Haller, 18. Unión esclero-coroidea.*

Los estudios acerca de la DMAE han sido realizados ampliamente en poblaciones de países nórdicos, americanos (del norte) y europeos, pero no en latinoamericanos. En este sentido, el criterio médico se basa en resultados de poblaciones que no necesariamente tienen las mismas condiciones clínicas, genéticas y/o sociales nuestras. Por esta razón, se hizo necesario generar una herramienta de apoyo a los especialistas oftalmólogos y retinólogos en el diagnóstico y pronóstico, y de apoyo a tratamientos tempranos de los pacientes con dicha patología aplicando algoritmos de machine learning [5,6].

En Colombia, existe falta de estudios que hayan abordado específicamente la estadificación de la atrofia geográfica en la DMAE utilizando la tecnología de OCT de alta definición. Esta tecnología proporciona imágenes más detalladas y precisas de la retina, lo que podría mejorar la capacidad de diagnóstico y seguimiento de la enfermedad [16]. Esto llevó a la necesidad de realizar una investigación enfocada en la estadificación imagenológica de la atrofia geográfica en la DMAE utilizando OCT de alta definición. Este estudio permitió evaluar la utilidad de la tecnología de OCT de alta definición en el diagnóstico y futuro seguimiento de la enfermedad [9, 10 y 12].

Puntualmente, el problema de ciencia de datos a abordar fue la falta de métodos exactos y eficientes para la estadificación de la atrofia geográfica en la Degeneración Macular Relacionada con la Edad (DMAE). La DMAE es una enfermedad ocular degenerativa que puede causar pérdida de la visión central y la atrofia geográfica es una de las etapas avanzadas de la enfermedad. El uso de la OCT con cinco cortes es una técnica de imagenología que proporciona información detallada sobre la estructura de la mácula en la retina y es una herramienta útil para la estadificación de la atrofia geográfica en la DMAE en el campo asistencial. No obstante, se hizo necesario desarrollar un enfoque específico para analizar y cuantificar los datos obtenidos de la OCT, con el fin de obtener una evaluación precisa y reproducible de la extensión y la gravedad de la atrofia geográfica.

El problema por afrontar fue, entonces, desarrollar un método de análisis de imágenes de OCT de alta definición para la estadificación precisa de la atrofia geográfica en la patología a estudio. Esto implicó la identificación de características específicas en las imágenes de OCT que estuvieran asociadas con la atrofia geográfica, el desarrollo de un modelo a partir de algoritmos de procesamiento de imágenes para cuantificar estas características y establecer de acuerdo con la literatura un sistema de clasificación para determinar la extensión y la gravedad de la atrofia geográfica con enfoque predictivo.

Se tuvo en cuenta, sin embargo, el afrontar algunas dificultades tales como el problema del ruido de las imágenes, en este caso en escala de grises, lo que exigió realización de algunas mejoras como el contraste. Se tuvo también el problema de la heterogeneidad de las imágenes, es decir, la variación significativa en las características, propiedades o contenido de las imágenes, lo que representó un desafío afectando el rendimiento de los algoritmos. Otras situaciones presentadas fueron el overfitting o sobreajuste y la variabilidad en el volumen de datos o imágenes por clase, lo que generaría un desbalance. Gracias al conjunto de datos que se logró adquirir (7321 imágenes), se pudo experimentar con diferentes arquitecturas buscando alcanzar un buen ajuste, utilizando herramientas como por ejemplo *data augmentation* para mejorar la capacidad de generalización del modelo al exponerlo a diferentes variaciones de los datos de entrada o dropout para mejorar la curva del error de la prueba regularizando y buscando prevenir el sobreajuste.

La expectativa en los resultados de este trabajo aplicado radicó en alcanzar un impacto significativo en el manejo clínico de la DMAE en pacientes del sur de Colombia, al proporcionar información importante sobre la progresión de la enfermedad y permitir una detección temprana y un potencial tratamiento oportunos. Además, se propone que este estudio siente las bases para futuras investigaciones en el campo de la ciencia de datos y la imagenología en el contexto de la degeneración macular.

1.2. FORMULACIÓN DEL PROBLEMA

¿Cómo desarrollar un método de análisis de imágenes de tomografía óptica coherente de segmento posterior de alta definición, para la estadificación precisa de la atrofia geográfica en la degeneración macular atrófica, relacionada con la edad?

El problema central de esta formulación es la falta de métodos automatizados y precisos para la estadificación de la atrofia geográfica en la degeneración macular seca utilizando imágenes de OCT de alta definición. Actualmente, la interpretación de estas imágenes depende en su totalidad de especialistas en retina altamente entrenados, lo que puede llegar a ser subjetivo o en cierta forma impreciso (*entre conceptos de especialistas*), consume tiempo y no siempre es accesible u oportuno, especialmente en regiones con recursos limitados.

El reto aquí consiste en desarrollar un método que automatice el análisis de las imágenes de tomografía óptica, identifique y extraiga características clave asociadas con los diferentes estadios de la atrofia geográfica, y clasifique con alta precisión estas imágenes, superando obstáculos técnicos como el ruido, la heterogeneidad de los datos y el desbalance de clases.

Abordar este problema permite mejorar significativamente el alcance diagnóstico y seguimiento de pacientes con DMAE, facilitando intervenciones más tempranas y efectivas. Además, se establece un método reproducible y escalable que puede implementarse en diversos entornos clínicos, contribuyendo al avance del conocimiento científico en oftalmología y ciencia de datos aplicada a la salud visual.

2. OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Desarrollar un modelo de aprendizaje automático que permita estadificar la degeneración macular atrófica geográfica relacionada con la edad (DMAE), a partir de los resultados de las imágenes de la tomografía óptica coherente (OCT) del segmento posterior del ojo en una población del sur de Colombia.

2.2 OBJETIVOS ESPECÍFICOS

i. Compilar un número importante de imágenes de OCT del segmento posterior del ojo de pacientes con degeneración macular no vascular (atrófica) relacionada con la edad.

ii. Implementar técnicas de aprendizaje profundo para analizar las imágenes recolectadas del OCT y desarrollar un modelo de clasificación imagenológico de la DMAE.

iii. Evaluar el desempeño del modelo predictivo basado en métricas establecidas en la literatura para clasificación.

3. MARCO DE REFERENCIA

3.1. MARCO TEÓRICO

3.1.1 De La Degeneración Macular Atrófica Relacionada Con La Edad:

La degeneración macular atrófica relacionada con la edad (DMAE) es una enfermedad que afecta a millones de personas en el mundo. Se calcula que alrededor de 200 millones de personas sufren algún tipo de DMAE a nivel global y alrededor de 10 millones en USA a 2021 [1]. Se espera que, para el año 2040 sean 300 millones de personas en el planeta [3]. En el mundo, la DMAE es la tercera causa de ceguera, después de catarata y glaucoma, y es más común después de los 55 años [2], aumentando progresivamente su riesgo a medida que aumenta la edad especialmente después de los 80 años. Existen dos tipos de DMAE: la degeneración macular neovascular o húmeda que afecta del 15% al 20% de la población con DMAE, pero que produce pérdida severa de la visión al 80% de los pacientes que la padecen. La otra degeneración macular es la no neovascular o seca, que afecta al 80% a 85% de la población y que generalmente tiene un pronóstico más favorable [1,7].

De acuerdo con Thomas et al en [1], La DMAE involucra cambios patológicos en las capas profundas de la retina, específicamente en la mácula y alrededor de la misma, generando pérdida de la visión central (Figura No. 5). Esta pérdida se da por acumulación de depósitos llamados drusas que son la piedra angular de los cambios imagenológicos de la patología [8]. La forma exudativa o húmeda de la DMAE, presenta formación de nuevos vasos (neovascularización) con crecimiento de membranas neovasculares coroideas que llevan a hemorragia y exudación entre las capas de la retina produciendo pérdida profunda de la visión (Figura No. 6).

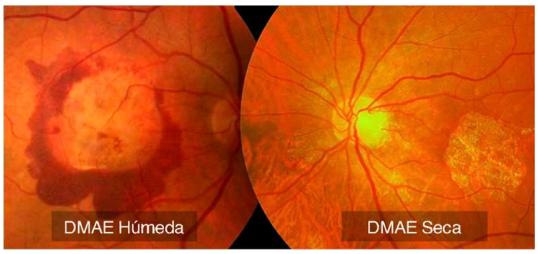
Estas membranas, son el resultado de una proliferación vascular anormal generadas por el factor de crecimiento vascular endotelial (VEGF por sus siglas en inglés), razón por la que para este tipo de DMAE el tratamiento resulta ser los anti-VEGF (antiangiogénicos), que son inyecciones intravítreas que impiden la acción del factor de crecimiento endotelial vascular [15].

Figura No. 5 – **Izquierda**: visión normal. **Derecha**: pérdida de la visión central.

En el caso de la DMAE seca o atrófica (no vascular), se genera una cicatriz atrófica o atrofia geográfica en la mácula, con lesiones que no responden al tratamiento con los anti-VEGF dado que, el problema aquí no es por el factor de crecimiento endotelial vascular [1,15].

En esta patología es importante tener en cuenta los factores de riesgo. Estos pueden ser de índole personal, sociodemográfico, incluso relacionados con el estatus socioeconómico [2]. Dentro de lo personal podemos mencionar la edad, el sexo, la raza/etnia, la genética [2]; a nivel ocular el color del iris, la densidad óptica del pigmento macular, la catarata (antecedente quirúrgico de catarata), error refractivo y el ratio entre la fóvea y el disco óptico; y factores personales sistémicos como la

enfermedad cardiovascular entre otros [2]. Existen también factores de riesgo para la progresión de la neovascularización coroidea como la presencia de cinco o más drusas, hiperpigmentación, hipertensión arterial, raza blanca y tabaquismo. Otros factores de riesgo son el índice de masa corporal alto, y la hiperlipidemia [1,8].



Los pacientes pueden no tener síntomas o tenerlos, pero mínimos, cuando la enfermedad se encuentra en su estadio más temprano, que generalmente, es cuando las drusas se encuentran en tamaño predominantemente pequeño o mediano.

Figura No. 6 – DMAE húmeda (**Izq**.), DMAE seca (**Der**).

Una vez los pacientes inician con la presentación de sintomatología, aluden metamorfopsia que es la distorsión visual, visión borrosa cercana especialmente con la lectura, disminución con la sensibilidad al contraste. En la DMAE húmeda, los síntomas se dan generalmente más rápido y con afectación más profunda de la visión, como severa distorsión, escotoma central (punto negro) (Figura No. 5) y dificultad para reconocer rostros, debido a hemorragia retiniana o acumulación de fluidos en la mácula (Figura No. 6). La DMAE geográfica, también presenta distorsión visual y escotoma central, pero su progresión es menos acelerada que la neovascular [1].

Anteriormente, el diagnóstico de la DMAE se realizaba basados en el examen clínico (*ej.: fondo de ojo*) al paciente o en la evaluación de fotografías del fondo de ojo a color [3]. En las últimas dos décadas, la tomografía óptica coherente (OCT) ha sido utilizada para la detección de las lesiones producidas por la degeneración macular, mejorando su resolución cada vez más [11,17]. La angiografía con fluoresceína continúa siendo útil para detectar la neovascularización coroidea y su localización y actividad [3]. La angiografía OCT emergió como una aproximación no invasiva.

La DMAE tiene varios sistemas de clasificación. Clasificación epidemiológica (Wisconsin grading) en donde se tipifica entre temprano y tardío; la clasificación clínica básica que incluye estadios intermedios y su relación con los cambios según la edad; y la clasificación según el estudio AREDS simplificada de escala de puntos de severidad. A continuación, en la Tabla No. 1, se evidencian las tres clasificaciones mencionadas [3] seguida de la clasificación de AREDS desde la revisión realizada [1] en la Tabla No. 2. Posteriormente, en la Tabla No. 3, encontraremos la clasificación más reciente y en la que se basó el presente estudio.

Tabla No. 1 – Definiciones y escalas de clasificación de la DMAE (*AMD – aged-related macular degeneration*)

	Definición
Clasificación Epidemiológica (Wisconsin Grading)	
AMD Temprana	Drusas grandes (> 125 µm) o pseudodrusas retinianas, o anormalidades pigmentarias
AMD Tardía	AMD neovascular o atrofia geográfica
Clasificación Clínica Básica *	
No cambios relacionados con la edad	No anormalidades pigmentarias ni drusas
Cambios normales relacionados con la edad	Solo drusas pequeñas ≤ 63 µm y sin anormalidades pigmentarias
AMD Temprana	Drusas de tamaño medio > 63 µm y ≤ 125 µm, y no anormalidades pigmentarias
AMD Intermedia	Drusas grandes > 125 µm o cualquier anormalidad pigmentaria
AMD Tardía	AMD neovascular o atrofia geográfica
Escala de puntos de severidad simplificada AREDS †	
0	No drusas grandes (> 125 µm) o cambios pigmentarios en cada ojo
1	Drusas grandes o cambios pigmentarios en un ojo solamente
2	Drusas grandes y cambios pigmentarios en un ojo solamente; o drusas grandes o cambios pigmentarios en ambos ojos; o AMD neovascular o atrofia geográfica en un ojo.
3	Drusas grandes y cambios pigmentarios en un ojo; y drusas grandes o cambios pigmentarios en el ojo contralateral
4	Drusas grandes y cambios pigmentarios en ambos ojos

***AMD** = degeneración macular relacionada con la edad. **AREDS** = estudio de enfermedades visuales relacionadas con la edad. * Las definiciones están basadas en el ojo más comprometido. † Un ojo con AMD tardía tiene un puntaje de 2.*

Fuente: Tomado de: Mitchell P. et al. Age-related macular degeneration. Lancet. 2018 Sep 29;392(10153):1147-1159. doi: 10.1016/S0140-6736(18)31550-2. PMID: 30303083. [3].

Tabla No. 2 – Clasificación AREDS

Clasificación de la degeneración macular relacionada con la edad basada en el estudio de enfermedades del ojo relacionadas con la edad.			
AMD Temprana	**AMD Intermedia**	**AMD No-neovascular Avanzada (AMD Seca Avanzada)**	**AMD Neovascular Avanzada (AMD Húmeda)**
Presencia de drusas pequeñas o algunas drusas de tamaño medio en uno o ambos ojos Cambios pigmentarios	Presencia de muchas drusas de tamaño medio, 1 drusa grande, y/o atrofia geográfica que no involucra la mácula central (fóvea)	Atrofia geográfica que involucra la mácula central o fóvea.	Neovascularización coroidea en 1 ojo.

Fuente: Tomado de: Thomas CJ, Mirza RG, Gill MK. Age-Related Macular Degeneration. Med Clin North Am. 2021 May;105(3):473-491. doi: 10.1016/j.mcna.2021.01.003. Epub 2021 Apr 2. PMID: 33926642. [1].

Tabla No. 3 – Clasificación de la atrofia en la degeneración macular relacionada con la edad por el consenso de la reunión de atrofia.

Clasificación de la atrofia en la degeneración macular relacionada con la edad por el consenso de la reunión de atrofia.	
Término	**Abreviación**
ERP completo y atrofia retinal externa	cRORA
ERP incompleto y atrofia retinal externa	iRORA
Atrofia retinal externa completa	cORA
Atrofia retinal externa incompleta	iORA

ERP = epitelio retinal pigmentario

Fuente: Tomado de: Sadda, S. R., Guymer, R., Holz, F. G., Schmitz-Valckenberg, S., Curcio, C. A., et al. (2018). *Consensus Definition for Atrophy Associated with Age-Related Macular Degeneration on OCT: Classification of Atrophy Report 3*. Ophthalmology, 125(4), 537-548. doi:10.1016/j.ophtha.2017.09.028.

3.1.1.1 De La Histopatología De La Retina En La Degeneración Macular Atrófica Relacionada Con La Edad:

Desde el ángulo histopatológico, la retina tiene 11 capas [1] que sirven de lecho para la presentación de las drusas [8]. Las drusas, son comprimidos de detritos celulares y lípidos entre otros, que se depositan en la membrana pigmentaria, que es la capa más externa de la retina, y ocurre progresiva y típicamente con la edad. Las drusas duras, de < 63 micras tienen bordes bien definidos y se pueden encontrar en toda la retina. Estas, no se asocian causalmente al desarrollo de la DMAE. Sin embargo, las drusas blandas que típicamente son > 125 micras de bordes definidos o no, sólo son encontradas en la mácula e incrementan significativamente el riesgo de la DMAE [4]. A continuación, se muestran las capas normales de la retina en Figura No. 7 y las formaciones de las drusas señaladas con flechas en la Figura No. 8.

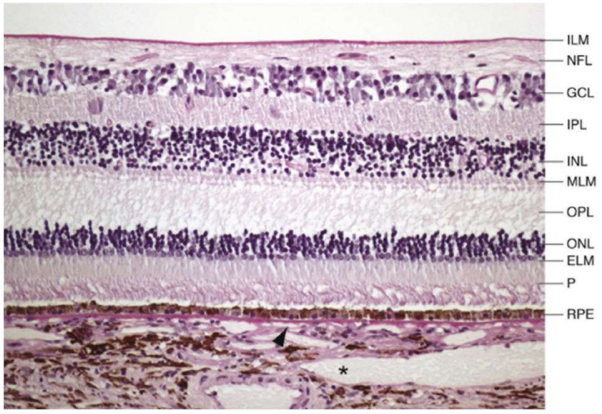


Figura No. 7 – Retina Normal. *Capas normales de la retina (tinción ácido peryódico de Schiff [PAS]). ELM, membrana limitante externa; GCL, capa de células ganglionares; ILM, membrana limitante interna; INL, capa nuclear interna; IPL, capa plexiforme interna; MLM, membrana limitante media; NFL, capa de fibras nerviosas; ONL, capa nuclear externa; OPL, capa plexiforme externa; P, fotorreceptores (segmentos internos/externos) de bastones y conos. Membrana de Bruch, cabeza de flecha; coroides. © Academia Americana de Oftalmología. [1].*

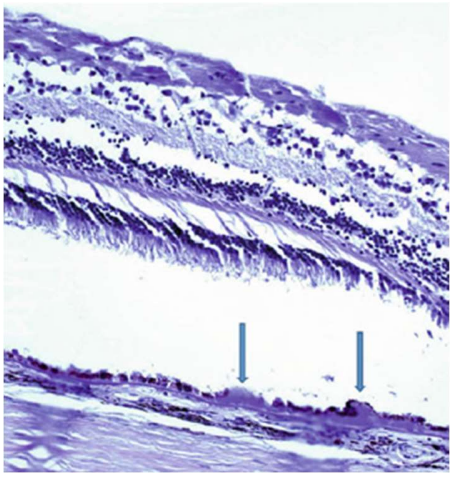


Figura No. 8 – Drusas. *Fotomicrografía de un ojo que muestra drusas (flechas) en la región macular (ácido peryódico de Schiff (PAS), aumento original, ×100). [4].*

3.1.1.2 De Las Imágenes Originadas En Tomógrafo De Coherencia Óptica En La Degeneración Macular Atrófica Relacionada Con La Edad:

Imagenológicamente, la tomografía óptica coherente (OCT) del segmento posterior del ojo, captura cortes que pueden ser horizontales o verticales visualizando las diferentes capas de la retina a nivel de la mácula y evidenciando cambios en la retina [9]. Estos cambios son principalmente las drusas depositadas en las capas más externas de la retina. A continuación, se muestra un corte normal (Figura No. 9) y un corte con presencia de drusas en la mácula (Figura No. 10) [1].

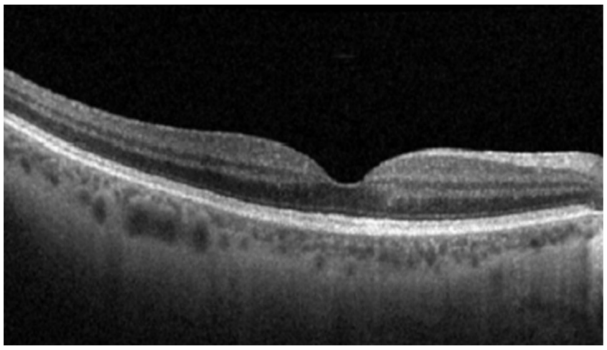


Figura No. 9 – OCT Normal. *OCT normal del ojo derecho que muestra capas retinianas intactas. (Imagen cortesía del Dr. M. Gill).*[1].

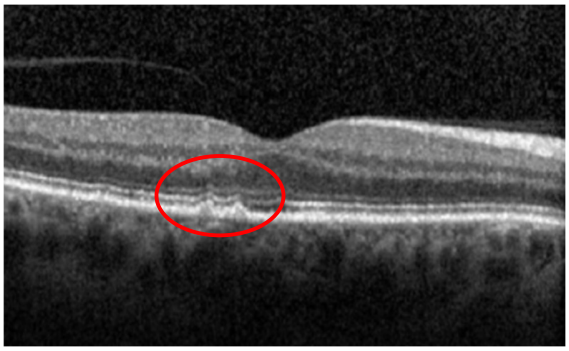


Figura No. 10 – Drusas en OCT. *DMAE seca con drusas maculares. (Imagen cortesía del Dr. M. Gill).* [1].

3.1.2 Del Aprendizaje Profundo Y Su Aplicabilidad En La Degeneración Macular Atrófica Relacionada Con La Edad:

En cuanto al aprendizaje profundo, esta se conoce como una subdisciplina del aprendizaje automático que utiliza redes neuronales artificiales con muchas capas, también conocidas como redes profundas, para modelar patrones complejos en los datos. A diferencia de los algoritmos de aprendizaje automático tradicionales, que requieren una significativa ingeniería de características, las redes profundas tienen la capacidad de aprender representaciones jerárquicas de datos de manera automática [24]. Estas redes neuronales profundas se entrenan utilizando grandes conjuntos de datos etiquetados y un algoritmo de optimización conocido como retropropagación. Este proceso ajusta los pesos de las conexiones neuronales con el objetivo de minimizar el error de predicción, permitiendo que las capas profundas de la red aprendan representaciones a diferentes niveles de abstracción, desde características de bajo nivel, como bordes en una imagen, hasta características de alto nivel, como formas y objetos completos [25] (Figura No. 11).

El aprendizaje profundo ha demostrado ser extremadamente eficaz en una amplia variedad de tareas, incluyendo el reconocimiento de patrones complejos, el reconocimiento de voz, la traducción automática, y, de manera crucial, el análisis de imágenes médicas. En el campo de la oftalmología, las redes neuronales profundas han mostrado su capacidad para aprender a detectar y clasificar enfermedades oculares a partir de imágenes, facilitando así el diagnóstico y tratamiento precoz [26].

Dentro del aprendizaje profundo, las redes neuronales convolucionales (CNN) destacan por su diseño específico para procesar datos con estructura de cuadrícula, como las imágenes. Estas redes utilizan operaciones de convolución para extraer características locales de los datos de entrada, lo que las hace especialmente adecuadas para tareas de visión por computadora [27]. Las CNN están compuestas por capas de convolución, capas de pooling y capas completamente conectadas. Las capas de convolución aplican filtros (kernels) para detectar características locales en las imágenes, mientras que las capas de pooling reducen la dimensionalidad de las características detectadas, preservando la información más relevante. Finalmente, las capas completamente conectadas integran la información extraída para realizar la clasificación [28].

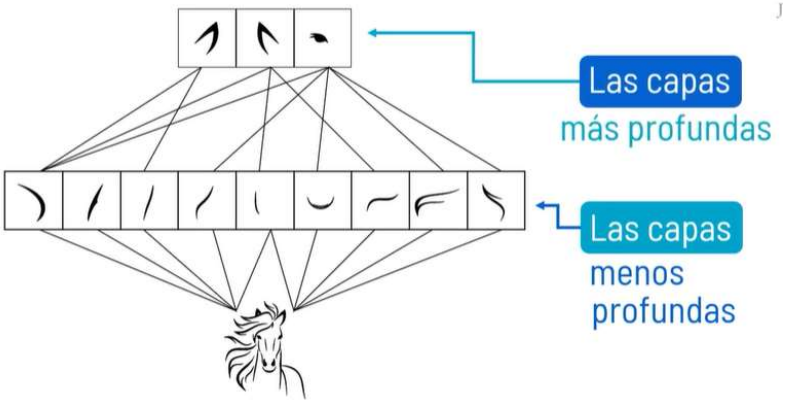


Figura No. 11 *– Aprendizaje de representaciones a diferentes niveles de abstracción de las imágenes.*

Extraído de asignatura 'Aprendizaje Profundo', Maestría en Ciencia de Datos – Universidad Javeriana Cali.

Debido a su capacidad para detectar patrones complejos y variaciones locales en los datos, las CNN se utilizan ampliamente en el análisis de imágenes médicas. Estas redes son empleadas en tareas como la detección de tumores, la segmentación de órganos, y el diagnóstico de enfermedades oculares a partir de imágenes obtenidas mediante OCT [29].

Las imágenes diagnósticas desempeñan un papel esencial en la medicina moderna, ya que permiten a los profesionales de la salud visualizar el interior del cuerpo humano y detectar anomalías de manera no invasiva. En el ámbito de la oftalmología, la OCT es particularmente crucial. Esta tecnología avanzada proporciona imágenes de alta resolución de la retina y la mácula, permitiendo a los médicos detectar y monitorizar enfermedades oculares como la degeneración macular y el glaucoma. Estas imágenes detalladas son fundamentales para el diagnóstico precoz y la gestión efectiva de patologías que pueden llevar a la pérdida de visión, lo que subraya su importancia en la atención oftalmológica tanto a nivel global como en regiones específicas como Colombia [30,31]. El uso del aprendizaje profundo en el análisis de imágenes diagnósticas ha revolucionado este campo, posibilitando el desarrollo de herramientas de diagnóstico automatizadas y precisas. Las CNN entrenadas en reconocer patrones específicos asociados con diversas enfermedades han mejorado la precisión del diagnóstico, además de reducir la carga de trabajo de los especialistas [32].

La tomografía óptica coherente es una técnica de imagen no invasiva que utiliza luz para capturar imágenes en alta resolución de las capas de la retina. Esta técnica es fundamental para el diagnóstico y seguimiento de enfermedades oculares, incluyendo la degeneración macular relacionada con la edad [33]. El proceso de adquisición de imágenes OCT implica el uso de un haz de luz dirigido al ojo. La luz reflejada por las diferentes capas de la retina es capturada y procesada para generar una imagen detallada de su estructura. Este método permite a los médicos observar cambios sutiles en la retina que pueden indicar la presencia y progresión de enfermedades [33,34].

La tecnología OCT tiene una importancia vital en la oftalmología a nivel mundial, debido a su capacidad para detectar enfermedades oculares en etapas tempranas. En Colombia, la adopción de estas tecnologías es crucial para mejorar la atención médica en regiones con acceso limitado a especialistas en retinología. El uso de modelos de aprendizaje profundo para analizar imágenes

OCT puede facilitar diagnósticos más rápidos y precisos, beneficiando a la población local [35].

La evaluación del desempeño de los modelos de aprendizaje profundo en el análisis de imágenes médicas se puede realizar mediante diversas métricas, como por ejemplo precisión, sensibilidad, especificidad, F1-Score y el área bajo la curva ROC (AUC-ROC), entre otras. Estas métricas permiten cuantificar la capacidad del modelo para identificar correctamente casos positivos y negativos, proporcionando una medida objetiva de su eficacia [36]. Una evaluación rigurosa es esencial para garantizar que los modelos de aprendizaje profundo sean fiables y seguros para su uso clínico. En el contexto de la oftalmología, la precisión del diagnóstico automatizado puede tener un impacto significativo en la salud visual de los pacientes, haciendo que la evaluación exhaustiva sea una prioridad [37].

3.1.2.1 De Las Estrategias De Aprendizaje De Máquinas y Aprendizaje Profundo, Y Métricas Aplicadas Al Proyecto:

En el contexto del presente proyecto de tesis de maestría sobre Redes Neuronales Convolucionales (CNN) para clasificación de imágenes médicas originadas en tomógrafo de coherencia óptica, existen varias estrategias de aprendizaje automático y aprendizaje profundo para optimizar el rendimiento de los modelos. Específicamente, arquitecturas como VGG16, ResNet y EfficientNet adicionales a una arquitectura propia son aplicables al diseño del trabajo. No obstante, también estrategias como extractores de características, junto con clasificadores como Support Vector Machines (SVM) y K-Nearest Neighbors (KNN-3), entre otros. A continuación, se detallan las estrategias y métricas que de acuerdo con la literatura pueden ser aplicadas para evaluar los modelos de acuerdo con los objetivos planteados.

3.1.2.1.1 Estrategias de Aprendizaje Automático:

Las arquitecturas de CNN preentrenadas—VGG16, ResNet y EfficientNet—que son utilizadas para extraer características de las imágenes en estos modelos, entrenados en el conjunto de datos ImageNet, capturan representaciones de características ricas que pueden transferirse a nuevas tareas [44,46,49,51]. Congelar las capas convolucionales de estas redes, puede generar ventajas con relación a sus mapeos de características aprendidos sin entrenamiento adicional, lo cual es una práctica común conocida como aprendizaje por transferencia [48].

Después de extraer características, se emplean clasificadores como SVM y KNN3 para la clasificación de imágenes. Las SVM son efectivas para datos de alta dimensionalidad y pueden encontrar el hiperplano óptimo que separa las clases [38]. El clasificador KNN es un método no paramétrico que clasifica instancias basándose en la etiqueta mayoritaria entre los k vecinos más cercanos en el espacio de características [39].

Se pueden explorar otros clasificadores de aprendizaje automático para mejorar la precisión y robustez del modelo. El algoritmo Random Forest es un método de ensamblado que construye múltiples árboles de decisión durante el entrenamiento y produce la clase que es el modo de las clases de los árboles individuales [55]. Es conocido por su capacidad para manejar datos de alta dimensionalidad y reducir el riesgo de sobreajuste. En este proyecto, Random Forest se plantea para clasificar las imágenes basándose en las características extraídas, proporcionando una interpretación más robusta y resistente al ruido de los datos.

XGBoost es una implementación optimizada de Gradient Boosting (GBM) que ha demostrado un rendimiento superior en diversas competiciones y aplicaciones [56]. GBM construye modelos predictivos de forma aditiva mediante la optimización de una función de pérdida diferenciable [57]. Ambos algoritmos se emplean para aprovechar su capacidad de manejar relaciones no lineales y capturar interacciones complejas entre características. Se ajustan los hiperparámetros para optimizar su rendimiento en el conjunto de datos de imágenes médicas.

Las redes neuronales feedforward (FFNN) son un tipo de red neuronal artificial donde la información se mueve en una sola dirección, desde las capas de entrada hasta las capas de salida, pasando por las capas ocultas [24]. Se implementan FFNN para clasificar las imágenes utilizando las características extraídas, permitiendo modelar relaciones no lineales y complejas en los datos. Se plantea experimentar con diferentes números de capas y neuronas para encontrar la arquitectura óptima.

Aunque las RNN están diseñadas principalmente para datos secuenciales, también se puede explorar su aplicación en este proyecto para capturar posibles dependencias espaciales en las características extraídas [58]. Si bien las imágenes estáticas no presentan una secuencia temporal, las RNN pueden ser útiles en el procesamiento de secuencias de características que podrían representar patrones relevantes para la clasificación.

Los métodos de ensamblado combinan múltiples modelos para mejorar el rendimiento general del sistema [59]. Se puede aplicar el método de *Ensemble Voting*, donde se combinan las predicciones de varios clasificadores (como SVM, KNN, Random Forest, XGBoost y FFNN) mediante votación mayoritaria o ponderada. Este enfoque ayuda a reducir la varianza y puede conducir a mejoras en la precisión y generalización del modelo.

Se cuenta con un gran conjunto de datos con un desbalance de clases, con algunas clases subrepresentadas, y una de ellas, sobrerrepresentada como es el caso de la clase 'Normal', se calculan pesos de clase inversamente proporcionales a las frecuencias de las clases y se aplican durante el entrenamiento del modelo. Este enfoque ayuda a mitigar el sesgo entre las diferentes clases [43]. En la clasificación SVM, se plantea la utilización del parámetro class_weight, mientras que en la KNN se plantea explorar técnicas de ponderación para ajustar el desbalance.

Se puede realizar técnica de validación cruzada k-fold estratificada, la cual asegura que cada partición del conjunto de datos preservará la misma distribución de clases que el conjunto original. Este método proporciona una estimación más confiable del rendimiento del modelo en conjuntos de datos desbalanceados [45].

Para evaluar cómo el rendimiento de los modelos escala con diferentes cantidades de datos de entrenamiento, se generan curvas de aprendizaje. Estas curvas ayudan a diagnosticar si los modelos se benefician de datos adicionales o si están subajustando o sobreajustando [42].

3.1.2.1.2 Estrategias de Aprendizaje Profundo:

El aprendizaje por transferencia permite aprovechar modelos preentrenados en grandes conjuntos de datos buscando mejorar el aprendizaje en una tarea objetivo con datos limitados [48]. Al utilizar VGG16 [49], ResNet [44] y EfficientNet [51], se toma ventaja de su capacidad para extraer características de alto nivel que resultan ser relevantes para el problema de clasificación

presentado en este trabajo de grado. Al realizar congelación de las capas convolucionales de los modelos preentrenados se evita la alteración de los pesos aprendidos durante el entrenamiento. Esta técnica reduce el costo computacional y previene el sobreajuste cuando el conjunto de datos es pequeño [52].

Otra estrategia es el preprocesamiento que incluye redimensionar las imágenes al tamaño de entrada requerido por cada arquitectura y normalizar los valores de los pixeles. También, se puede plantear la aplicación de ecualización de histograma para mejorar el contraste de las imágenes, lo que podría mejorar la extracción de características [41]. En teoría, implementar entrenamiento de precisión mixta mixed_float16 permitiría un cómputo más rápido y reducción del uso de memoria sin pérdida significativa de precisión del modelo [47].

3.1.2.1.3 Métricas Planteadas para la Evaluación del Modelo:

3.1.2.1.3.1 Métricas del Reporte de Clasificación:

i. **Precisión:** Mide la corrección de las predicciones positivas, calculada como verdaderos positivos divididos entre la suma de verdaderos y falsos positivos [50].

ii. **Recall (Sensibilidad):** Mide la capacidad del modelo para identificar todas las instancias relevantes, calculada como verdaderos positivos divididos entre la suma de verdaderos positivos y falsos negativos [50].

iii. **F1-Score:** Media armónica de la precisión y el recall, proporcionando un equilibrio entre ambos [50].

iv. **Soporte:** Número de ocurrencias reales de cada clase en el conjunto de datos.

3.1.2.1.3.2 Exactitud:

La exactitud general calcula la proporción de instancias clasificadas correctamente sobre el total de instancias [40]. Sin embargo, en conjuntos de datos desbalanceados, la exactitud puede ser engañosa, por lo que se complementa con otras métricas.

3.1.2.1.3.3 Matriz de Confusión:

La matriz de confusión proporciona un desglose detallado de las clasificaciones correctas e incorrectas para cada clase, permitiendo la identificación de patrones específicos de error de clasificación [40].

3.1.2.1.3.4 Curvas ROC y Área Bajo la Curva (AUC):

Las curvas ROC trazan la tasa de verdaderos positivos contra la tasa de falsos positivos en varios umbrales. El AUC representa la probabilidad de que el clasificador asigne una puntuación más alta a una instancia positiva elegida al azar que a una negativa elegida al azar [40].

3.1.2.1.3.5 Especificidad:

La especificidad mide la proporción de verdaderos negativos identificados correctamente, importante para evaluar el rendimiento del modelo en clases negativas [50].

3.1.2.1.3.6 Curvas de Aprendizaje:

Las curvas de aprendizaje trazan el rendimiento del modelo en los conjuntos de entrenamiento y validación contra el tamaño del conjunto de entrenamiento. Son útiles para diagnosticar problemas de sesgo y varianza [42].

Al plantear la integración de clasificadores de aprendizaje automático como SVM y KNN con características extraídas de modelos de aprendizaje profundo, se busca combinar las fortalezas de ambos paradigmas. Es de tener en cuenta que, en el transcurso de las pruebas se buscará la aplicación de otros clasificadores. El uso de aprendizaje por transferencia con arquitecturas como VGG16, ResNet y EfficientNet permitirá una extracción eficiente de características de las imágenes, aprovechando el amplio entrenamiento que estos modelos pueden experimentar en grandes conjuntos de datos [46, 49, 44, 51].

El manejo del desbalance de clases será crucial debido a la representación desigual de las clases en el conjunto de datos. Aplicar ponderación de clases y validación cruzada estratificada aseguraría que los clasificadores no se sesgaran hacia las clases mayoritarias [43, 45].

Las métricas de evaluación elegidas proporcionarán una evaluación integral de los modelos. Confiar únicamente en la exactitud podría resultar engañoso, por lo que, incorporar precisión, recall, F1-score, especificidad, curvas ROC y AUC brindaría una comprensión más matizada del rendimiento del modelo, especialmente en presencia del actual desbalance de clases [40, 50].

3.2. ANTECEDENTES Y TRABAJOS RELACIONADOS

Existen trabajos realizados con relación al tema que se desarrolla en este estudio, sin embargo, la población estudiada ha sido predominantemente europea, nórdica o de américa del norte. Se hace, por tanto, necesario estudiar el comportamiento de la patología degenerativa macular en nuestra población. A continuación, se presentan dos estudios que materializan la intención del presente trabajo.

El primero se titula "Clinically relevant deep learning for detection and quantification of geographic atrophy from optical coherence tomography: a model development and external validation study" (*Aprendizaje profundo clínicamente relevante para la detección y cuantificación de la atrofia geográfica de la tomografía óptica coherente: un estudio de modelo de desarrollo y validación externa*), realizado en 2021, el cual utiliza cuatro diferentes modelos de aprendizaje profundo para la segmentación de la atrofia geográfica y sus características. Para el modelo, se segmentó el Dataset de manera manual de 200 pacientes (399 ojos) con atrofia geográfica secundaria a DMAE. El desempeño se validó externamente con un Dataset de pacientes reclutados externamente. Se buscaba obtener un consenso entre la opinión de dos expertos técnicos y el modelo. Dicha validación se realizó en 110 pacientes (192 ojos). Los resultados del modelo fueron similares al consenso de los especialistas humanos. Finaliza sugiriendo que la aplicación del modelo en la práctica de rutina aportaría información prometedora para el diagnóstico y pronóstico de pacientes en validaciones futuras. [5,7].

El segundo estudio, realizado en 2022, titula" Prediction of visual function from automatically quantified optical coherence tomography biomarkers in patients with geographic atrophy using machine learning" (*Predicción de la función visual a partir de biomarcadores cuantificados automáticamente en la tomografía óptica coherente en pacientes con atrofia geográfica usando aprendizaje de máquinas*). Este es un estudio no intervencionista de análisis de pacientes con atrofia geográfica post hoc secundaria a la DMAE no neovascular. Los datos de imágenes fueron recolectados de dos fuentes, una de pacientes del estudio FILLY trial (NCT02503332) y de las atenciones reales de pacientes no incluidos en estudios. A todos los pacientes se les realizó sólo un OCT de mácula por ojo. Todas las imágenes fueron procesadas utilizando modelos de aprendizaje profundo. Los modelos fueron entrenados para cada una de las características que definen la atrofia geográfica (pérdida de la membrana pigmentaria, degeneración de los fotorreceptores e hipertransmisión). Un cuarto modelo que segmenta el tipo RORA de la atrofia también fue entrenado para definir si la lesión era completa (cRORA) o incompleta (iRORA). Adicionalmente, se utilizó un modelo de regresión tipo random forest que fue entrenado usando los resultados segmentados. Fueron 476 ojos (325 pacientes) con atrofia geográfica secundaria a DMAE seca. La edad media de los pacientes en el estudio era de 80.5 años. Los resultados del análisis demostraron que la agudeza visual de pacientes con atrofia geográfica secundaria a DMAE seca se podía predecir usando las características del OCT que fueron automáticamente segmentadas y cuantificadas. El algoritmo pudo producir imágenes en mapa de calor generando información simultánea de la localización de las lesiones en el OCT. Los resultados demostraron la

utilidad de los biomarcadores de imágenes segmentadas automáticamente en la predicción de la función visual, contribuye al desarrollo de la toma de decisiones más precisas y personalizadas por paciente. [6,8].

Estos trabajos evidencian el aporte al estudio de la DMAE en cuanto a su aproximación diagnóstica y pronostica, y deja puertas abiertas a la continuidad que se debe hacer a estos análisis para lograr mayor precisión en el manejo de estos pacientes. Adicionalmente, aportan en la generación de información útil para nuevos estudios enfocados en los posibles tratamientos [18]. La idea con estos ejemplos fue poder realizar un estudio en ciencia de datos que nos permitiera alcanzar este nivel de precisión, pero con población nuestra, aportando a la información potencialmente requerida en los nuevos estudios de tratamientos para la patología en cuestión [19].

4. PREPARACIÓN Y DESARROLLO DEL PROYECTO

Durante la realización del presente trabajo de grado, se llevaron a cabo múltiples pruebas utilizando una base de datos de 7,321 imágenes obtenidas del tomógrafo de coherencia óptica de alta definición de la Clínica Visual y Auditiva del Instituto para Niños Ciegos y Sordos. Estas imágenes, recopiladas entre los años 2020 y 2024 de una población del sur de Colombia, proporcionaron el sustrato fundamental para este estudio. A continuación, se presenta el desarrollo del trabajo según los pasos planteados en los objetivos específicos.

4.1. PREPARACIÓN Y COMPILACIÓN DE LAS IMÁGENES DE TOMOGRAFÍA ÓPTICA COHERENTE (OCT).

> *Objetivo No. 1 – Compilar un número importante de imágenes de tomografía óptica coherente (OCT) del segmento posterior del ojo de pacientes con degeneración macular no vascular (atrófica) relacionada con la edad*

En la búsqueda de alcanzar el objetivo de desarrollar un modelo preciso para la estadificación de la degeneración macular atrófica relacionada con la edad (DMAE) utilizando imágenes de tomografía óptica coherente (OCT), el primer objetivo específico se centró en compilar un número significativo de imágenes OCT del segmento posterior del ojo de pacientes con DMAE no vascular (atrófica). Este paso fundamental implicó la recolección de imágenes de alta calidad (HD) que representaran adecuadamente la población afectada y cumplieran con criterios estrictos para su análisis. Para lograrlo, se trabajó en varias actividades que incluyeron la obtención de aprobación ética, la identificación y selección de imágenes que cumplieran con los criterios de inclusión establecidos, la organización sistemática y almacenamiento seguro de los datos, y la exploración inicial de las imágenes utilizando herramientas especializadas de procesamiento, preparándolas así para etapas posteriores del estudio.

4.1.1 ESTRUCTURA DE LA REVISIÓN SISTEMÁTICA DE LA LITERATURA

Se llevó a cabo entonces, la revisión de la literatura científica y médica existente sobre la degeneración macular atrófica relacionada con la edad y su diagnóstico mediante tomografía de coherencia óptica. Además, se exploraron estudios que integran técnicas de aprendizaje automático y profundo en el análisis de imágenes de OCT, así como el uso de métricas adecuadas para evaluar los resultados.

La revisión se realizó consultando diferentes bases de datos como PubMed, IEEE Xplore, ScienceDirect y Scopus, abarcando publicaciones predominantemente entre 2013 y 2023 para garantizar la actualidad de la información. Se utilizaron combinaciones de palabras clave y términos MeSH relacionados con la DMAE, OCT, aprendizaje automático, aprendizaje profundo y métricas de evaluación, incluyendo términos como "Age-Related Macular Degeneration" (*degeneración macular relacionada con la edad*), "Geographic Atrophy" (*atrofia geográfica*),

"Optical Coherence Tomography" (*tomografía de coherencia óptica*), "Machine Learning" (*aprendizaje de máquinas*) y "Metrics" (*métricas*). Los criterios de inclusión consideraron algunos artículos revisados por pares, otros por revisión propia, disponibles en texto completo, en inglés o español, que abordaran aspectos clínicos de la DMAE atrófica, su diagnóstico mediante OCT, aplicaciones de aprendizaje automático y profundo, y el uso de métricas de evaluación en imágenes médicas que estuvieran dentro del contexto del proyecto. Se excluyeron estudios no relacionados directamente con la DMAE atrófica, centrados exclusivamente en la forma exudativa, publicaciones duplicadas o resúmenes sin datos completos.

El proceso de selección inició con la identificación de alrededor de 150 artículos potencialmente relevantes para el estudio. Tras eliminar duplicados y revisar títulos y resúmenes, se evaluaron 80 textos completos para verificar el cumplimiento de los criterios de inclusión. Finalmente, y a lo largo del desarrollo del trabajo de grado, se seleccionaron 59 referencias que aportaban información significativa a los objetivos del estudio. Se extrajo información clave de cada artículo y se agruparon según temas principales: aspectos clínicos de la degeneración macular atrófica, uso de OCT en el diagnóstico, aplicaciones de aprendizaje automático y profundo, y métricas de evaluación empleadas.

Las métricas comunes para evaluar el rendimiento de los modelos incluyen precisión, sensibilidad, especificidad, puntuación F1 (*F1-Score*) y área bajo la curva ROC (AUC) [35,39,49]. La elección de métricas adecuadas fue crucial, especialmente en conjuntos de datos desbalanceados [42]. Fawcett [35] y Sokolova y Lapalme [49] proporcionan guías para la interpretación y aplicación de estas métricas.

De acuerdo con la literatura revisada, se puede inferir que la combinación de tomografía óptica con técnicas de aprendizaje automático y específicamente profundo mejora significativamente el diagnóstico y seguimiento de la DMAE atrófica. Se evidenció en esta revisión como los modelos desarrollados permiten no solo la detección automática de lesiones, sino también la cuantificación y predicción de la progresión de la enfermedad [5,10,16] según la clasificación empleada. Sin embargo, se identifican desafíos como la necesidad de grandes conjuntos de datos etiquetados, con un adecuado manejo del desbalance de clases y con la interpretación clínica de los resultados proporcionados por los modelos [9,42] como se pudo evidenciar también en el desarrollo del presente documento. Además, la variabilidad en las técnicas y métricas empleadas dificulta la comparación directa entre estudios, resaltando la importancia de estandarizar metodologías [35,49].

4.1.2 OBTENCIÓN DE APROBACIÓN ÉTICA Y CONSIDERACIONES ÉTICAS

El presente trabajo de grado se presentó ante el Comité de Ética e Investigación del Instituto para Niños Ciegos y Sordos, que se reúne mensualmente, donde fue aprobado. Este comité recibe propuestas, avances y resultados de nuevos estudios de investigación realizados con apoyo

institucional y en colaboración con diferentes universidades con las que tiene convenios, los cuales impactan la misión y visión de esta entidad sin ánimo de lucro.

El objetivo principal de este comité es garantizar la protección de los derechos y el bienestar de los participantes en estudios científicos (pacientes), evaluando la validez y el rigor de los protocolos antes de su aprobación. Supervisa el desarrollo de los estudios, monitoreando su conformidad con las normativas éticas y científicas, y gestiona adecuadamente cualquier evento adverso. Además, fomenta la transparencia, evita conflictos de interés y capacita a los investigadores en ética, facilitando la comunicación con la comunidad científica y el público para promover la confianza en el proceso de investigación.

4.1.3 DISEÑO DE LOS CRITERIOS DE INCLUSIÓN Y SELECCIÓN DE IMÁGENES

En colaboración con el director del proyecto, PhD H. Vargas, y la codirectora, Dra. D. Libreros, se decidió recolectar un mínimo de 800 imágenes por clase de pacientes con DMAE, basándose en la "*clasificación de la atrofia en la degeneración macular relacionada con la edad por el consenso de la reunión de atrofia*" y aplicando criterios de inclusión específicos. Estos criterios consideraban pacientes del sur de Colombia, con edades entre 50 y 99 años, cuyos estudios se hubieran realizado entre los años 2020 y 2024. Las imágenes debían ser obtenidas mediante exámenes de OCT realizados con el equipo Zeiss Cirrus 5000 HD de la Clínica Visual y Auditiva del Instituto para Niños Ciegos y Sordos.

Criterios de Inclusión para recolección de datos para Trabajo de Grado DMAE		
1.	Población:	Sur de Colombia
2.	Rango de edad:	Entre 50 y 99 años
3.	Rango de años de recolección de la data:	2020 a 2024
4.	Equipo biomédico utilizado para obtención de imágenes:	Tomógrafo de Coherencia Óptica - OCT marca Zeiss Cirrus 5000 HD
5.	Tipo de imágenes a recolectar:	Estudios de OCT normales y con criterio clínico de DMAE, de ambos ojos siempre que la patología ocurra de manera bilateral
6.	Ubicación anatómica de imágenes de OCT:	Mácula
7.	Color de las imágenes:	Blanco y negro
8.	Tipo de scan:	HD 5 Line Raster' y 'HD 5 Line'
9.	Fuerza de la señal (1 a 10):	≥ 6
10.	Ángulo del scan:	0° (*cortes horizontales*)
11.	Longitud de la imagen:	Entre 6 y 9 mm
12.	Número de cortes:	3 cortes
13.	Cortes utilizados:	2, 3, y 4
14.	Formato de imagen:	.png

Fuente: elaboración propia

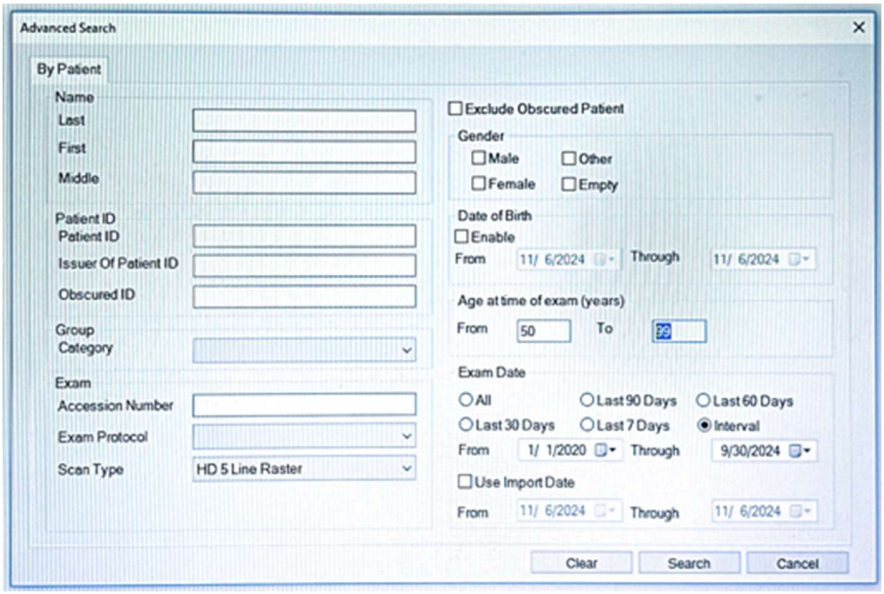
Tabla No. 4 – Criterios de inclusión de los datos. *Se definieron 14 criterios de selección para la recolección de los datos para el presente trabajo de grado.*

Se incluyeron tanto estudios de OCT normales como aquellos con diagnóstico clínico de DMAE atrófica, considerando imágenes de ambos ojos siempre que la patología fuera bilateral. Las imágenes seleccionadas correspondían a la región de la mácula, debían ser en escala de grises y haberse obtenido mediante los tipos de escaneo "HD 5 Line Raster" o "HD 5 Line" en el equipo

biomédico. Se exigió una fuerza de señal igual o superior a 6 en una escala de 1 a 10, y el ángulo de escaneo debía ser de 0°, es decir, cortes horizontales. La longitud de las imágenes oscilaría entre 6 y 9 mm de longitud.

Figura No. 12 – Búsqueda avanzada en OCT. *Visual de la búsqueda avanzada en el equipo biomédico OCT Cirrus 5000 HD de Zeiss. Aquí se selecciona el tipo de scan (Scan Type), el rango de edades (Age at a time of exam - years), y el rango de fechas (Exam Date) para la búsqueda de las imágenes.*

Fotografía tomada a la pantalla del equipo.



Aunque el OCT realiza cinco cortes numerados del 1 al 5, se seleccionaron específicamente los cortes 2, 3 y 4, ya que son los cortes más centrales y que mejor representan la mácula; por lo tanto, se descartaron los cortes 1 y 5. Todas las imágenes se almacenaron en formato .png para facilitar su posterior análisis.

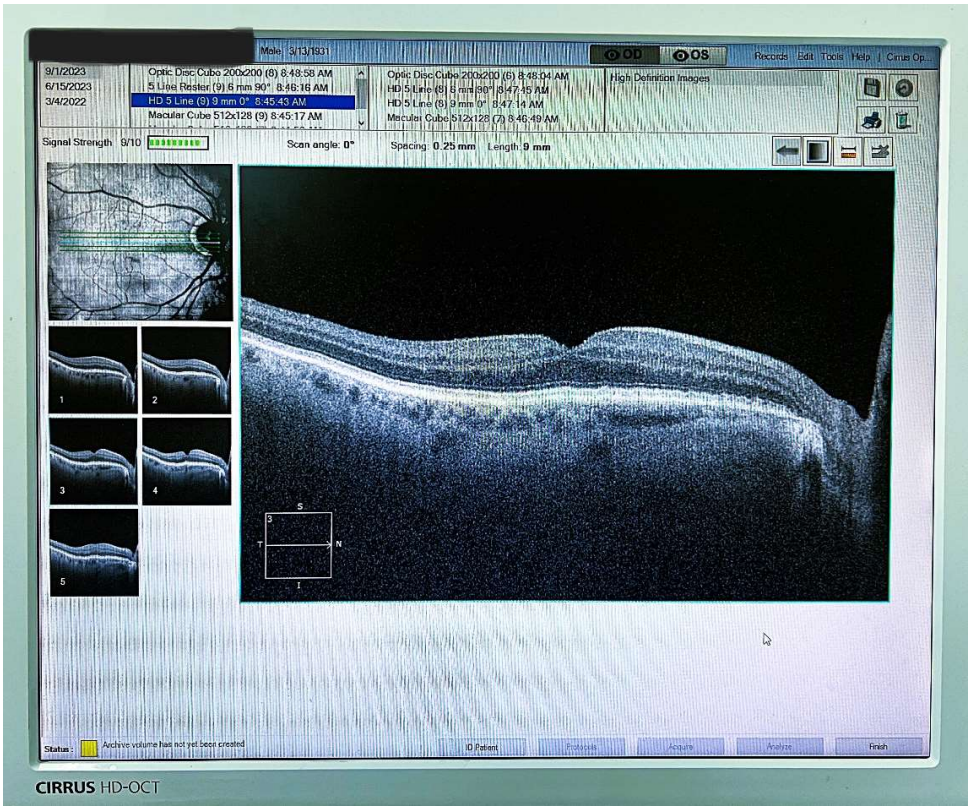


Figura No. 13 – Visual de selección de las imágenes. *Imagen que evidencia cómo se realizó el proceso de selección de las imágenes. Se observa también que, con cada imagen se pueden validar algunos de los criterios de selección establecidos en el proyecto.*

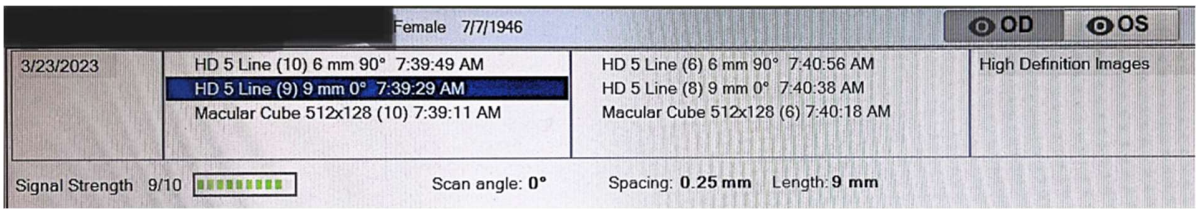


Figura No. 14 – Parámetros de selección de las imágenes. *Parte superior de la pantalla en donde se validaron los criterios de selección.*

A pesar de no alcanzar el número inicialmente planeado de imágenes por clase, se trabajó con las imágenes disponibles que cumplían con los criterios de inclusión establecidos. Al finalizar la recolección, se obtuvieron 4,521 imágenes para la clase "Normal", 646 imágenes para "iORA" (Atrofia Retiniana Externa Incompleta), 545 imágenes para "cORA" (Atrofia Retiniana Externa Completa), 680 imágenes para "iRORA" (Atrofia Incompleta del Epitelio Pigmentario Retiniano y Retiniana Externa) y 929 imágenes para "cRORA" (Atrofia Completa del Epitelio Pigmentario Retiniano y Retiniana Externa). En total, se recopilaron 7,321 imágenes para el estudio (ver Tabla no. 5).

Tabla No. 5 – Inventario final de datos recolectados.

Inventario FINAL	--> en el Drive de: aquinonesmd2					
	Normal	iORA	cORA	iRORA	cRORA	Total
Total	4.521	646	545	680	929	7.321
Faltaron --> para llegar a 800		-154	-255	-120	129	-400 faltaron en total
		-19%	-32%	-15%	16%	

Fuente: elaboración propia

4.1.4 ORGANIZACIÓN Y ALMACENAMIENTO DE LAS IMÁGENES

Se estableció un sistema de almacenamiento utilizando Google Drive, donde se creó una carpeta principal denominada "DMAE". Dentro de esta carpeta, s e generaron cinco subcarpetas correspondientes a cada una de las clases trabajadas: "Normal", "iORA", "cORA", "iRORA" y "cRORA". En total, se recopilaron 7,321 imágenes de OCT, con un tamaño total de 4.88 GB.

Para acceder y manipular estas imágenes desde Google Colab, se montó Google Drive en el entorno de Python. Primero, se importó el módulo ‘drive’ de ‘google.colab’, que permite interactuar con Google Drive desde Colab. Luego, se utilizó ‘drive.mount('/content/drive')’ para conectar Google Drive al entorno de Colab, solicitando autorización al usuario. Una vez montado, se asignó la ruta del directorio de imágenes a la variable ‘original_dataset_dir’ (/content/drive/My Drive/DMAE/…), facilitando el acceso a los archivos. Con este paso se pudo acceder y manipular los datos almacenados en Google Drive dentro de Google Colab.

4.1.5 EXPLORACIÓN INICIAL DE LAS IMÁGENES RECOLECTADAS

Se procedió a analizar las imágenes de OCT recopiladas, conforme a los objetivos previamente establecidos, con el fin de caracterizar la base de datos según la patología en estudio—la degeneración macular seca con atrofia geográfica—con enfoque en el análisis de clasificación. La recolección de las imágenes se realizó de manera consistente, teniendo en cuenta parámetros de captura uniformes según lo establecido en los criterios de inclusión, utilizando el mismo equipo OCT del Instituto para Niños Ciegos y Sordos, y clasificando las imágenes según la "*clasificación de la atrofia en la degeneración macular relacionada con la edad por el consenso de la reunión de atrofia*".

Se llevó a cabo una curación del conjunto de datos, validando la calidad de las imágenes y descartando aquellas que no cumplían con los criterios establecidos o que presentaban borrosidad o artefactos. Las imágenes se descargaron en formato .png, asegurando una resolución uniforme con una fuerza de señal igual o superior a 6.

El etiquetado fue validado por la retinóloga codirectora del proyecto, garantizando la correcta clasificación de las imágenes. Se realizó un análisis exploratorio inicial para comprender mejor las características del conjunto de datos, creando cuadrículas de imágenes representativas de cada clase para detectar patrones visuales y revisando la distribución del número de imágenes por clase para identificar posibles desbalances.

Posteriormente, se aplicó un preprocesamiento a las imágenes para normalizarlas y facilitar su comparación. Se implementó data augmentation con el objetivo de aumentar la diversidad del conjunto de datos sin introducir sesgos.

4.2. DISEÑO E IMPLEMENTACIÓN DEL MODELO DE CLASIFICACIÓN IMAGENOLÓGICA

> *Objetivo No. 2 – Implementar técnicas de aprendizaje de máquinas para analizar las imágenes recolectadas del OCT y desarrollar un modelo de clasificación imagenológico de la DMAE.*

Para avanzar en el diagnóstico preciso y oportuno de la Degeneración Macular Atrófica Relacionada con la Edad, fue fundamental aprovechar las ventajas que ofrecen las técnicas modernas de aprendizaje automático y profundo. El objetivo específico número 2 se centró en implementar técnicas de aprendizaje de máquinas para analizar las imágenes recolectadas del tomógrafo y desarrollar un modelo de clasificación imagenológico de la DMAE. Este objetivo busca no solo automatizar el proceso de detección y clasificación de la atrofia geográfica en las imágenes de tomografía de coherencia óptica, sino también mejorar la precisión y eficiencia en la identificación de las distintas etapas de la enfermedad.

4.2.1 ETIQUETADO Y CLASIFICACIÓN DE LAS IMÁGENES

Se llevó a cabo el etiquetado de las 7,321 imágenes de OCT, clasificándolas en cinco carpetas o clases según la clasificación consensuada de la atrofia en la degeneración macular relacionada con la edad: "Normal", "iORA", "cORA", "iRORA" y "cRORA" (ver Tabla No. 3).

Se consideró la posibilidad de aplicar métodos de segmentación de imágenes para extraer las áreas afectadas, sin embargo, tras una evaluación detallada, se determinó que, en el contexto de este proyecto centrado en la clasificación supervisada utilizando redes neuronales convolucionales (CNN) y modelos robustos de aprendizaje profundo, no era estrictamente necesario implementar técnicas de segmentación para alcanzar los resultados esperados.

El enfoque adoptado se basó en el uso de modelos preentrenados y clasificadores eficientes, como por ejemplo EfficientNet, capaces de extraer automáticamente características relevantes de las imágenes. Estas redes neuronales pueden aprender a identificar patrones y características distintivas asociadas con cada clase, incluyendo las regiones de atrofia geográfica, sin necesidad de una segmentación previa explícita [5,16]. Esto permitió lograr representaciones suficientemente ricas para llevar a cabo una clasificación lo más exacta posible. La segmentación es una técnica valiosa cuando se requiere aislar características anatómicas específicas dentro de las imágenes, como diferentes capas de la retina o lesiones particulares. Sin embargo, su implementación habría incrementado significativamente la complejidad computacional y los recursos necesarios [23,31,43,48] implicando mayores tiempos de procesamiento de los modelos, lo que podría no haber sido eficiente para el propósito principal del proyecto. Además, al confiar

en las capacidades de las CNN para enfocarse en las áreas más relevantes durante el proceso de aprendizaje, se optimizó el rendimiento sin agregar pasos adicionales. No obstante, se aseguró un etiquetado consistente de todas las imágenes, lo cual fue fundamental para el aprendizaje supervisado y para que los modelos pudieran diferenciar correctamente entre las distintas clases. Este proceso de etiquetado detallado permitió que las redes neuronales identificaran las características clave asociadas con la atrofia geográfica en las imágenes de OCT. Los resultados obtenidos son consistentes con los objetivos planteados y demuestran que la segmentación no era imperativa. Empero, para futuras etapas del proyecto, si se busca mejorar aún más la exactitud o analizar características más detalladas dentro de este tipo de imágenes, la segmentación podría ser una técnica complementaria valiosa a considerar.

4.2.2 APLICACIÓN DE TÉCNICAS DE PREPROCESAMIENTO

Se buscó aplicar técnicas de preprocesamiento a las imágenes de tomografía óptica posterior para mejorar su calidad y resaltar características relevantes. Se implementaron diversas operaciones en el proyecto para optimizar el rendimiento del modelo de clasificación. Estas técnicas incluyeron la normalización, ecualización del histograma, corrección de iluminación, eliminación de artefactos, redimensionamiento de imágenes y aumento de datos.

Inicialmente, se realizó la normalización de las imágenes para estandarizar los valores de intensidad de los píxeles a una escala uniforme entre 0 y 1. Este proceso apuntó a reducir el impacto de variaciones en la iluminación y el contraste entre las imágenes, facilitando la interpretación por parte del modelo y mejorando la consistencia de los datos de entrada. Según González y Woods (2002) [40], la normalización es una práctica común en el procesamiento digital de imágenes para asegurar que las diferencias en intensidad no influyan negativamente en el análisis. Además, se aplicó la ecualización del histograma mediante la función cv2.equalizeHist() de OpenCV. Esta técnica redistribuye los valores de intensidad de los píxeles para mejorar el contraste global de la imagen, resaltando las estructuras anatómicas clave en las imágenes OCT. Esto es necesario para la estadificación precisa de las diferentes etapas de la atrofia geográfica en la degeneración macular. González y Woods (2002) [40] igualmente destacan que la ecualización del histograma es efectiva para resaltar detalles en imágenes con contrastes reducidos. La ecualización del histograma también contribuyó a la corrección de iluminación, abordando inconsistencias en el brillo y la exposición de las imágenes. Al uniformar la iluminación, se aseguró que el modelo no fuera influenciado por variaciones no deseadas que pudieran afectar su capacidad de generalización. Este paso garantizó que las características relevantes fueran las que guiaran el proceso de aprendizaje del modelo, tal como se menciona en Goodfellow et al. (2016) [41].

En cuanto a la eliminación de artefactos y reducción de ruido, aunque no se aplicaron filtros avanzados explícitos como filtros de mediana o gaussianos, se incorporaron técnicas de filtrado implícito al trabajar con métodos de redimensionamiento y normalización. Estas operaciones ayudaron a eliminar artefactos y reducir el ruido de baja intensidad, mejorando la calidad general

de las imágenes seleccionadas y facilitando la detección de patrones significativos por parte del modelo [41]. Las imágenes fueron redimensionadas para adaptarse a los requisitos de entrada de las arquitecturas de redes neuronales utilizadas, como ResNet o EfficientNet. Este paso aseguró una coherencia en las dimensiones de las imágenes de entrada, optimizando el procesamiento y reduciendo el costo computacional. Tan y Le (2019) [50] plantean que EfficientNet logra un equilibrio entre exactitud y eficiencia computacional, lo cual es beneficioso para manejar grandes conjuntos de datos de imágenes médicas.

Para incrementar la diversidad del conjunto de datos y prevenir el sobreajuste al que se estaba expuesto dada la base de datos con la que se contó, se implementaron técnicas de aumento de datos (*data augmentation*). Estas incluyeron transformaciones como rotaciones, volteos horizontales y verticales, desplazamientos y escalados. Al generar variaciones de las imágenes existentes, se mejoró la capacidad del modelo para generalizar a nuevos datos no vistos, lo que fue fundamental para este proyecto de aprendizaje profundo [41].

Adicionalmente, se realizaron ajustes en los formatos y canales de color de las imágenes para asegurar la compatibilidad con las expectativas del modelo. Dado que las imágenes eran en escala de grises, se adaptaron para que el modelo pudiera procesarlas correctamente, ya que algunas arquitecturas preentrenadas esperan entradas con tres canales (RGB). González y Woods (2002) [40] enfatizan la importancia de adaptar el formato de las imágenes a los requisitos del sistema para evitar incompatibilidades y pérdidas de información.

El uso de técnicas de entrenamiento de precisión mixta también contribuyó a mejorar la eficiencia computacional del proyecto. Al implementar la política mixed_float16, se pudo acelerar el entrenamiento y reducir el uso de memoria sin afectar significativamente la precisión del modelo. Micikevicius et al. (2018) [46] demostraron que el entrenamiento de precisión mixta permite aprovechar al máximo las capacidades del hardware moderno, optimizando el rendimiento sin sacrificar exactitud.

4.2.3 UTILIZACIÓN DE TÉCNICAS DE EXTRACCIÓN DE CARACTERÍSTICAS

En concordancia con la búsqueda de técnicas de extracción de características para capturar aspectos relevantes de las imágenes de tomografía de coherencia óptica aplicando descriptores específicos como histogramas de intensidad, texturas o características basadas en formas, se implementaron enfoques automatizados basados en redes neuronales convolucionales. Las CNN son conocidas por su capacidad para realizar extracción automática de características a través de sus capas convolucionales [23,24,26]. Estas capas utilizan filtros entrenables que identifican patrones importantes en las imágenes, como bordes, texturas y cambios estructurales característicos de las diferentes etapas de la degeneración macular atrófica [27,31]. El modelo implementado aprovecha las propiedades de las CNN para aprender representaciones jerárquicas de las imágenes tomográficas estudiadas. En las primeras capas, los filtros convolucionales detectan características básicas como bordes y contornos, mientras que en capas más profundas

capturan características más complejas y específicas de la patología [24,25]. Este enfoque permite que el modelo identifique automáticamente las áreas afectadas por la atrofia geográfica sin necesidad de definir manualmente descriptores como histogramas de intensidad o texturas (Figura No. 11).

Adicionalmente, el uso de técnicas de preprocesamiento como la normalización y la ecualización del histograma contribuyó a mejorar la calidad de las imágenes y, por ende, la eficacia en la extracción de características [40]. La normalización garantiza que los valores de intensidad de los píxeles estén en una escala consistente, lo que facilita el aprendizaje del modelo al reducir la variabilidad no deseada [41]. Por su parte, la ecualización del histograma mejora el contraste de las imágenes, resaltando las estructuras anatómicas relevantes y permitiendo que los filtros convolucionales detecten más fácilmente las características esenciales para la clasificación [40].

Al utilizar arquitecturas preentrenadas como ResNet, EfficientNet, y VGG16, se aprovechó el aprendizaje por transferencia, donde los modelos previamente entrenados en grandes conjuntos de datos pueden aplicar su conocimiento a una nueva tarea [47,50]. Esto buscó permitir la extracción de características de alto nivel de las imágenes OCT sin requerir un entrenamiento desde cero, tratando de optimizar el rendimiento y buscando reducir el costo computacional [50] en las pruebas realizadas, pese al tamaño de la base de datos del proyecto.

4.2.4 SELECCIÓN Y AJUSTE DEL MODELO PREDICTIVO ADECUADO

Se buscó seleccionar el modelo predictivo adecuado para el problema de clasificación imagenológica de la degeneración macular atrófica, para lo que se implementaron y evaluaron diversos algoritmos de aprendizaje automático para la clasificación. El objetivo era identificar el modelo que ofreciera el mejor rendimiento en la clasificación de imágenes de tomografía óptica para la estadificación precisa de la DMAE.

Inicialmente, se aplicó el modelo K-Nearest Neighbors (KNN) con k=3, un clasificador basado en la vecindad que predice la clase de una nueva instancia en función de las tres instancias más cercanas en el espacio de características [38]. KNN utilizó las características extraídas previamente por diversas redes neuronales convolucionales (CNN), como la propia desarrollada en el proyecto, ResNet, EfficientNet y VGG16. Aunque KNN es simple y efectivo en ciertos contextos, se encontró que su desempeño no era óptimo debido a la demanda computacional asociada con grandes volúmenes de datos y la alta dimensionalidad de las características, específicamente en las arquitecturas ResNet y EfficientNet, lo que es consistente con las limitaciones mencionadas por Cover y Hart (1967) [38].

Posteriormente, se exploró el uso de Máquinas de Vectores de Soporte (*Support Vector Machines* – SVM) con kernel de base radial (RBF), que busca encontrar el hiperplano óptimo que separa las clases con el máximo margen posible [37]. SVM es eficaz en espacios de alta dimensionalidad y puede manejar problemas no lineales. En este proyecto, SVM utilizó las características extraídas de la capa de salida de la CNN propia desarrollada, permitiendo capturar patrones complejos

presentes en las imágenes tomográficas. Los resultados mostraron una mejora significativa en la exactitud de clasificación (Gráfico No. 1).

También se implementó el algoritmo de Random Forest, un modelo de ensamblado que combina múltiples árboles de decisión para mejorar la robustez y capacidad de generalización [55]. Random Forest fue aplicado utilizando las características extraídas por las CNN y se basó en la votación mayoritaria de múltiples árboles para predecir la clase de cada imagen. Este modelo maneja bien datos desbalanceados y reduce el riesgo de sobreajuste, tal como lo destaca Breiman (2001) [55], pero puede ser menos eficiente con un gran número de árboles y presentar mayores tiempos de entrenamiento afectando la eficiencia del modelo.

Además, se empleó XGBoost, una implementación eficiente de árboles de decisión potenciados (gradient boosting), que construye modelos de forma secuencial optimizando los errores de los modelos anteriores [56]. XGBoost mostró una alta exactitud y eficiencia computacional gracias a su paralelismo y regularización, sobretodo aplicada a la CNN propia, aunque el ajuste de sus hiperparámetros puede ser complejo y sensible al ruido en los datos, según lo indicado por Chen y Guestrin (2016) [56].

Se consideró también el uso de Gradient Boosting Machine (GBM), otra técnica de potenciación que ajusta modelos secuencialmente para mejorar el rendimiento general [57]. GBM es efectivo en datos complejos y puede capturar relaciones no lineales, pero puede ser más lento y propenso al sobreajuste en comparación con XGBoost, como señala Friedman (2001) [57].

En el ámbito de las redes neuronales, se implementó una un Multilayer Perceptron (*Perceptrón multicapa* – MLPClassifier feed forward), que incluye varias capas densas y es capaz de capturar relaciones no lineales en los datos [41]. Esta red fue entrenada con las características extraídas por las CNN, permitiendo modelar patrones complejos presentes en las imágenes OCT. Sin embargo, requiere tiempo de entrenamiento para evitar el sobreajuste, según las observaciones de Goodfellow et al. (2016) [41].

Asimismo, se exploró el uso de una Red Neuronal Recurrente (RNN) con una capa SimpleRNN para capturar posibles patrones secuenciales en las características extraídas [58]. Aunque las RNN son ideales para datos secuenciales, su aplicación en imágenes estáticas es menos directa. En este proyecto, las RNN no mostraron mejoras significativas en el rendimiento para esta tarea específica comparado con los otros modelos, lo que coincide con las limitaciones mencionadas por Lipton et al. (2015) [58].

Finalmente, se implementó un Ensamblado de Modelos mediante Ensemble Voting, combinando los modelos con mejores resultados SVM, FFNN y RNN para mejorar la robustez y exactitud de la clasificación [59]. Este enfoque aprovecha las fortalezas individuales de cada modelo, reduciendo la varianza y los errores individuales al tomar una decisión conjunta basada en la votación. Opitz y Maclin (1999) [59] destacan que los métodos de ensamblado pueden mejorar significativamente el rendimiento en comparación con modelos individuales.

Después de realizar múltiples pruebas y comparaciones, se determinó que el modelo SVM aplicado a las características extraídas por la CNN propia desarrollada específicamente para este proyecto alcanzó la mejor exactitud (accuracy), con un valor promedio de 87.80%. Este resultado resalta la eficacia de combinar un extractor de características personalizado con un clasificador potente como SVM, configurado para el conjunto de datos y el problema específico de la clasificación de la DMAE.

4.2.5 SEPARACIÓN DEL CONJUNTO DE DATOS EN ENTRENAMIENTO Y VALIDACIÓN

Para este fin se empleó un esquema *hold-out* aleatorio. Esta estrategia aseguró una evaluación robusta y confiable del rendimiento de los modelos, alineándose con las mejores prácticas en aprendizaje automático y validación de modelos [44,41].

Utilizando el módulo *ImageDataGenerator* de Keras [36], el conjunto de datos se dividió en un 80% para *entrenamiento* y un 20% para *validación*. El conjunto de entrenamiento se empleó para ajustar las redes neuronales convolucionales utilizadas [27,50], mientras que el conjunto de validación sirvió para monitorear el rendimiento del modelo durante el entrenamiento y ajustar los hiperparámetros, ayudando a prevenir el sobreajuste [44,41].

En este proyecto, se decidió no utilizar un conjunto de prueba separado debido a consideraciones de disponibilidad de datos. Al destinar el 80% de los datos al entrenamiento y el 20% a la validación, se tuvo en cuenta el uso de los recursos disponibles, reduciendo la carga computacional y el tiempo necesario para completar los experimentos. Esto permitió que el modelo tuviera acceso a una cantidad suficiente de datos para aprender patrones significativos, mientras que el conjunto de validación proporcionó una estimación confiable de su rendimiento. Esta decisión se alineó con las necesidades prácticas del proyecto y las limitaciones en recursos computacionales, asegurando un balance entre la exactitud del modelo y la eficiencia operativa.

El motivo principal fue desarrollar y ajustar un modelo de clasificación eficaz para la patología retiniana utilizando las imágenes disponibles. La validación continua durante el entrenamiento permitió ajustar los hiperparámetros y monitorear el rendimiento sin necesidad de un conjunto de prueba separado [41]. Esta práctica es común en proyectos exploratorios o cuando se dispone de conjuntos de datos limitados, ya que una división adicional podría reducir demasiado el tamaño de los conjuntos de entrenamiento y validación, afectando negativamente el aprendizaje del modelo [42].

Se reconoce la importancia de evaluar el modelo en un conjunto de prueba independiente para obtener una medida más robusta de su capacidad de generalización. Por ello, en futuras etapas del proyecto se planteará recopilar más datos para implementar una evaluación con un conjunto de prueba separado y mejorar la capacidad computacional, fortaleciendo así la confiabilidad de los resultados.

4.3. EVALUACIÓN Y OPTIMIZACIÓN DEL MODELO PREDICTIVO

Objetivo No. 3 – Evaluar el desempeño del modelo predictivo basado en métricas establecidas en la literatura para clasificación.

Después de desarrollar el modelo predictivo para la clasificación imagenológica de la DMAE, se debe evaluar su desempeño para medir exactitud, precisión y confiabilidad. El objetivo número 3, se centra en esta evaluación, basada en métricas establecidas en la literatura para clasificación. Este objetivo busca medir la eficacia del modelo en la identificación correcta de las diferentes etapas de la atrofia geográfica en imágenes OCT, y también garantizar que el modelo sea confiable y aplicable en situaciones del mundo real. Las actividades asociadas abordan la aplicación de técnicas de validación y métricas de evaluación apropiadas, la realización de ajustes y mejoras adicionales cuando el rendimiento no es satisfactorio, y la ejecución de pruebas exhaustivas para verificar el correcto funcionamiento y validez del sistema implementado. Lo anterior, permite confirmar que el modelo predictivo no solo funciona correctamente desde una perspectiva técnica, sino que también cumple con los estándares necesarios para contribuir de manera efectiva al diagnóstico y seguimiento de la degeneración atrófica.

4.3.1 EVALUACIÓN UTILIZANDO MÉTRICAS APROPIADAS

En el presente proyecto, se desarrollaron varios modelos predictivos para clasificar las etapas de la atrofia geográfica asociada a la patología. Para evaluar el rendimiento de estos modelos, se utilizaron métricas de evaluación apropiadas para el contexto de soporte al diagnóstico clínico.

Debido al tamaño significativo de nuestro conjunto de datos—7,321 imágenes con un peso total de 4.88 GB—se enfrentaron retos computacionales que influenciaron la estrategia de evaluación. Aunque inicialmente se consideró la implementación de técnicas de validación cruzada k-fold para una evaluación más exhaustiva, el costo computacional asociado al procesamiento repetido de un volumen de datos tan grande lo hizo impracticable [42]. La validación cruzada requiere múltiples ciclos de entrenamiento y validación, lo cual habría incrementado significativamente el tiempo de procesamiento y el uso de recursos.

En lugar de ello, se optó por una partición del conjunto de datos en 80% para entrenamiento y 20% para validación como se mencionó anteriormente, de manera aleatoria (*Random hold-out*), haciendo 5 repeticiones (training-validation), y se reportan los valores promedio con su respectiva desviación estándar para cada métrica. Esta estrategia permitió maximizar la cantidad de datos disponibles para el entrenamiento del modelo, asegurando al mismo tiempo una estimación confiable de su rendimiento a través del conjunto de validación [44]. El conjunto de entrenamiento fue utilizado para ajustar los parámetros del modelo, mientras que el conjunto de validación sirvió para monitorear su desempeño y realizar ajustes necesarios para prevenir el sobreajuste [41].

Para la evaluación de los modelos de aprendizaje automático, se emplearon métricas estándar ampliamente reconocidas en la literatura de clasificación, tales como exactitud (*accuracy*), sensibilidad (*recall*), especificidad, F1-score y AUC-ROC [49]. La exactitud se calculó como el porcentaje de predicciones correctas respecto al total de predicciones realizadas, proporcionando una medida general del rendimiento del modelo. Además, se generó una matriz de confusión que permitió analizar detalladamente el rendimiento del modelo en cada clase específica, identificando verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos [40].

Para evaluar la capacidad discriminativa de los modelos entre las diferentes clases, se calcularon las Curvas Características Operativas del Receptor (*ROC*) y el Área Bajo la Curva (*AUC*) para cada clase, así como un AUC promedio macro [35,49]. Las curvas ROC ofrecen una representación visual del compromiso entre la sensibilidad y la especificidad del modelo a diferentes umbrales, mientras que el AUC proporciona una medida agregada de su capacidad para distinguir entre clases.

Dado el desbalance de clases en el presente estudio, y común en conjuntos de datos médicos [42], también se utilizaron el F1-score ponderado, que considera tanto la precisión como la recuperación, y pondera el impacto de cada clase según su soporte [49]. Esta métrica es particularmente útil para obtener una evaluación equilibrada del rendimiento del modelo cuando algunas clases están sub-representadas.

Adicionalmente, se calcularon la sensibilidad (recall) y la especificidad para cada clase y a nivel general. Estos indicadores son relevantes en contextos médicos, donde los costos de los falsos negativos y falsos positivos pueden ser significativamente diferentes [49].

4.3.2 AJUSTES Y MEJORAS DE LOS MODELOS

Durante el desarrollo del proyecto, surgieron diversos retos, principalmente derivados del tamaño y peso del conjunto de datos que generaron un alto costo computacional y afectaron el rendimiento de los modelos en diferentes etapas. En respuesta a estos desafíos, se realizaron ajustes y mejoras adicionales, que incluyeron la modificación de hiperparámetros, la incorporación de técnicas de regularización y la exploración de diferentes arquitecturas de red.

La fase experimental involucró el uso de diversos modelos basados en redes convolucionales y la afinación de sus hiperparámetros. Para las CNN preentrenadas como VGG16 [48] y EfficientNetB0 [50], se estableció una tasa de aprendizaje de 1e-3 utilizando el optimizador Adam [41], tamaños de lote (*batch size*) de 32, y un máximo de 50 épocas de entrenamiento, implementando técnicas de detención temprana (early stopping) para evitar sobre-entrenamiento.

En los modelos de aprendizaje automático clásicos, como K-Nearest Neighbors (KNN) [38], Support Vector Machines (SVM) [37], Random Forest [55] y XGBoost [56], se ajustaron hiperparámetros específicos: se utilizó k=3 para KNN, un kernel 'rbf' para SVM, y se establecieron 100 árboles para Random Forest y XGBoost. Estos ajustes permitieron incrementar el rendimiento de cada modelo acorde a las características de los datos y la complejidad del problema.

Para prevenir el sobreajuste y mejorar la capacidad de generalización, se aplicaron técnicas de regularización. En las redes neuronales, se aplicó Dropout [41] con una tasa que varió entre 0.4 y 0.6 en las capas densas, reduciendo la dependencia en neuronas específicas durante el entrenamiento. Además, se congelaron las capas base de los modelos preentrenados, manteniendo sus pesos para aprovechar las características ya aprendidas en grandes conjuntos de datos [47]. Se calcularon pesos por clase utilizando compute_class_weight de Scikit-Learn [36] para compensar el desbalance de clases presente en los datos, mejorando así la sensibilidad en las clases menos representadas [42].

Se emplearon también callbacks como EarlyStopping, ReduceLROnPlateau y ModelCheckpoint, útiles para regularizar el proceso de entrenamiento, prevenir el sobreajuste y ajustar dinámicamente la tasa de aprendizaje [41]. Estas técnicas permitieron detener el entrenamiento cuando no se observaban mejoras significativas en el correr de las épocas y reducir la tasa de aprendizaje en respuesta al estancamiento en la mejora del rendimiento.

Adicionalmente, se aplicaron modelos clásicos de aprendizaje automático sobre las características extraídas con *Deep Learning*, como SVM, Random Forest y XGBoost, MLP, MLPClassifier (Feedforward) con capas densas de 64 y 32 neuronas, y se implementó una Red Neuronal Recurrente utilizando una capa SimpleRNN con 32 unidades. Por último, se implementó un modelo de ensamblado (Ensemble Voting), que combinó SVM, Feedforward MLP y RNN utilizando una votación de tipo soft, buscando aprovechar las fortalezas de cada clasificador [59].

4.3.3 DOCUMENTACIÓN DEL PROCESO METODOLÓGICO Y PRESENTACIÓN DE RESULTADOS

Se documentó detalladamente todo el proceso metodológico y el desarrollo del proyecto, incluyendo cada paso seguido, los conjuntos de datos utilizados, las técnicas aplicadas y los resultados obtenidos. Esta documentación abarcó desde la recopilación y preprocesamiento de las imágenes de Tomografía Óptica, describiendo procedimientos como la normalización y la ecualización del histograma para mejorar la calidad de las imágenes. Se especificaron las arquitecturas de red exploradas, incluyendo VGG16 y EfficientNetB0, y se detallaron los modelos de aprendizaje automático implementados, como Support Vector Machines (SVM), Random Forest y XGBoost. Se registraron los ajustes de hiperparámetros realizados y las técnicas de regularización aplicadas, como Dropout, Early Stopping y el uso de pesos por clase para abordar el desbalance de datos.

4.4 SELECCIÓN DE MODELOS CNN PARA EXTRACCIÓN DE CARACTERÍSTICAS

4.4.1 EVALUACIÓN DE ESTRUCTURAS DE MODELOS – PRUEBAS PRELIMINARES Y AJUSTES PROGRESIVOS

Se realizaron algunas pruebas, que resultaron en las primeras tres versiones del código "*Estadif_DMAE_A_Quiñones*" representando fases preliminares de experimentación en el proyecto. Estas versiones fueron desarrolladas para explorar diversas estrategias y enfoques de modelado con un subconjunto limitado de datos, lo que permitió evaluar metodologías iniciales antes de utilizar el conjunto de datos completo.

La versión 1, se centró en pruebas preliminares utilizando una base de datos limitada con un muestreo de 230 imágenes por clase. Aquí se implementaron estrategias de balanceo de datos, análisis exploratorio y transferencia de aprendizaje utilizando VGG16. Esta arquitectura pre-entrenada se empleó para aprovechar los pesos aprendidos en conjuntos masivos como ImageNet. Sin embargo, debido al tamaño reducido del subconjunto de datos y la falta de ajuste de hiperparámetros, el rendimiento del modelo mostró resultados limitados, con una exactitud de validación del 46%. Esta versión sirvió para observar la viabilidad inicial del uso de arquitecturas preentrenadas y para establecer un punto de partida en la experimentación. En la versión 2, se aumentó el número de imágenes a 400 por clase, mejorando el balanceo del conjunto de datos. Se optó por una arquitectura simple de CNN, con tres capas convolucionales y técnicas de regularización como Dropout. Esta versión exploró arquitectura propia en lugar de utilizar modelos preentrenados. Aunque el modelo mostró mejoras en la capacidad de generalización, seguía siendo una fase exploratoria con datos preliminares, destinada a probar la simplicidad y efectividad del diseño de la arquitectura en tareas de clasificación de imágenes médicas. La versión 3, avanzó hacia una mejor versión del modelo mediante el uso de técnicas más sofisticadas, como la precisión mixta (*mixed precision*), que mejoró la eficiencia computacional y redujo los tiempos de entrenamiento. Se incorporó normalización batch y otras técnicas de regularización, como EarlyStopping y ReduceLROnPlateau. Lo anterior, logró un ajuste más fino de los hiperparámetros y mejoró la convergencia del modelo, reduciendo el riesgo de sobreajuste en esta prueba. Además, la evaluación se llevó a cabo utilizando un conjunto más amplio de métricas (precisión, recall, F1-Score y curvas AUC-ROC), proporcionando una evaluación integral del rendimiento.

Las tres versiones mencionadas fueron iniciales para la exploración del problema. Sin embargo, debido al uso de un subconjunto reducido de datos, no fueron consideradas para los experimentos finales del proyecto, ya que su propósito principal era realizar pruebas preliminares y sentar las bases para el desarrollo y ajuste de los modelos definitivos con la base de datos completa.

4.4.2 ANÁLISIS COMPARATIVO DE LAS VERSIONES 4 Y 5, Y SELECCIÓN DEL MODELO DEFINITIVO BASADO EN RESULTADOS DE EXACTITUD Y CAPACIDAD DE GENERALIZACIÓN

Posteriormente, al contar con la base de datos completa con las 7.321 imágenes, se implementaron las versiones 4 y 5. El análisis comparativo de las versiones 5, 5_1 y 5_2 muestra una clara evolución en el desarrollo del modelo para la clasificación de imágenes de OCT enfocadas en la estadificación de la DMAE. Estas versiones, aunque no fueron seleccionadas como modelos finales en el proyecto, representan una fase de experimentación importante que proporcionó aprendizajes valiosos para el refinamiento del modelo definitivo.

La versión 5: "*Estadif_DMAE_A_Quiñones_v5.ipynb*" implementa un enfoque donde se aumentó el número de imágenes a 600 por clase y se eliminaron los subconjuntos intermedios, lo que permitió trabajar con un conjunto de datos más extenso y representativo. El uso de precisión mixta (FP16) mejora la eficiencia computacional y reduce el uso de memoria durante el entrenamiento. La arquitectura de la CNN se optimiza con cinco capas convolucionales, incluyendo Batch Normalization, regularización L2, y Dropout. En la etapa de entrenamiento, se aplican estrategias como Data Augmentation, EarlyStopping, y ReduceLROnPlateau, lo que incrementa la variabilidad de los datos y ajusta dinámicamente el learning rate para mejorar la convergencia del modelo. Esta versión ofrece un análisis de las métricas de evaluación, proporcionando una base sólida para futuras iteraciones, aunque no alcanzó el rendimiento necesario para ser considerada como modelo base final. La versión 5_1: "*Estadif_DMAE_A_Quiñones_v5_1.ipynb*", refina aún más la estrategia aplicada en la versión anterior, incrementando las variaciones del Data Augmentation y simplificando el flujo de trabajo al eliminar el uso de subconjuntos. Este cambio permite un enfoque directo en el conjunto de datos original, mejorando la representatividad y simplificando el proceso de entrenamiento. Además, se ajusta la tasa de Dropout al 50%. Se implementan ajustes dinámicos en la tasa de aprendizaje mediante un Learning Rate Scheduler. Esta versión muestra mejoras significativas en el rendimiento del modelo. La versión 5_2: "*Estadif_DMAE_A_Quiñones_v5_2.ipynb*", es la iteración más avanzada de esta serie, integrando ajustes basados en los aprendizajes previos y empleando imágenes de mayor tamaño (384x384 píxeles). La arquitectura CNN se mantiene con precisión mixta y cinco capas convolucionales con regularización L2. El uso de Data Augmentation continúa siendo una estrategia clave. Además, se sigue usando EarlyStopping, ReduceLROnPlateau, y ajustes dinámicos del learning rate mediante un Learning Rate Scheduler. La evaluación que incluye la matriz de confusión muestra mejoras en la capacidad de generalización. Estas versiones se utilizaron como pruebas preliminares para explorar diferentes configuraciones y estrategias. Aunque ofrecieron mejoras significativas en comparación con versiones iniciales, no lograron superar el desempeño obtenido en la versión 4, que mostró una mayor exactitud y potencial para la estadificación imagenológica de la DMAE.

Comparando las versiones 4 y 5, encontramos que la versión 4 utiliza una arquitectura de CNN propia. La serie de versiones 5, incluyendo *Estadif_DMAE_A_Quiñones_v5*, *v5_1*, y *v5_2*,

representan iteraciones en el desarrollo de modelos para el presente proyecto de degeneración atrófica macular. Estas versiones implementaron progresos relevantes en técnicas de preprocesamiento, aumento de datos, regularización y optimización del entrenamiento, pese a ello, no lograron superar el rendimiento obtenido por la versión 4: "*DMAE_A_Quiñones_Original_v4.ipynb*", que alcanzó una exactitud del 71%, destacándose como el mejor modelo base hasta ese momento. A diferencia de la serie 5, la versión 4 utiliza una arquitectura de CNN propia. Aunque las versiones 5 introdujeron estrategias más avanzadas, como el uso de precisión mixta (FP16), regularización L2 más intensa, Dropout incrementado, y un Data Augmentation más variado, estas modificaciones no se tradujeron en mejoras significativas en la exactitud y, en algunos casos, afectaron negativamente la capacidad del modelo para generalizar ante nuevos datos. En contraste, la versión 4 aplicó un aumento de datos moderado y técnicas para mejora del entrenamiento menos complejas, pero logró un desempeño superior, mostrando mayor consistencia en la clasificación de las diferentes clases de DMAE. Además, la estabilidad del proceso de aprendizaje en la versión 4 permitió alcanzar resultados comparables sin recurrir a ajustes dinámicos complejos de la tasa de aprendizaje como los implementados en las versiones posteriores. Aunque las versiones 5 presentaron evaluaciones exhaustivas mediante métricas detalladas y técnicas avanzadas de ajuste, no lograron superar el 0.71 de exactitud obtenido en la versión 4, lo que indica que las mejoras introducidas no aportaron un rendimiento superior. Por estas razones, se decidió seleccionar la versión 4: "*DMAE_A_Quiñones_Original_v4.ipynb*" como el modelo definitivo de extracción de características para el proyecto, al demostrar un mejor equilibrio entre simplicidad, robustez y capacidad de generalización, así como un mayor potencial de mejora con ajustes futuros, consolidándola como la opción más eficiente para la estadificación imagenológica de la DMAE.

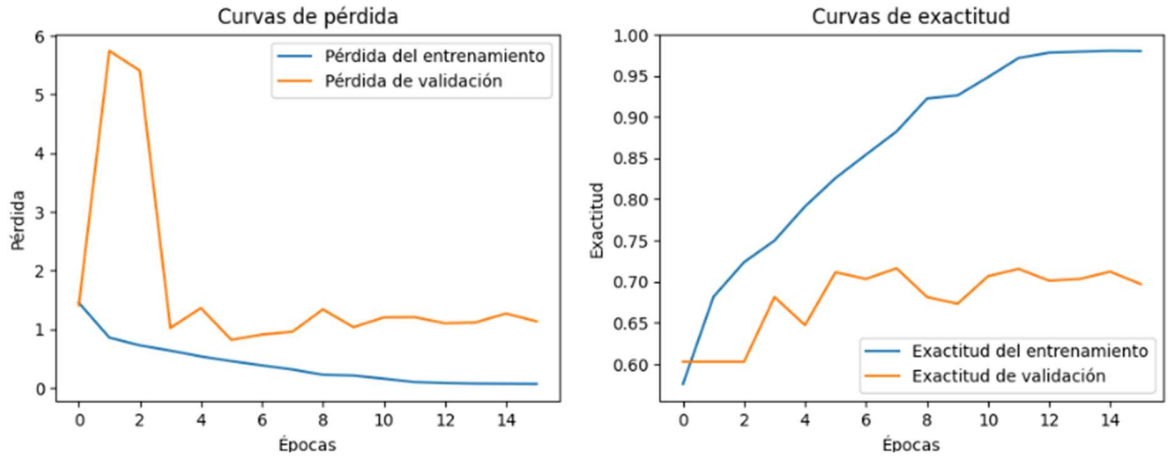


Figura No. 15 – Curvas de pérdida y exactitud *de la versión No. 4 de la fase inicial experimental del proyecto.*

4.4.3 CNN (Extracción de características) + APRENDIZAJE AUTOMÁTICO (Clasificación)

Durante el desarrollo del presente proyecto, y posterior a la evaluación de las versiones preliminares y selección del modelo 'base' (versión 4), se implementaron y evaluaron consecutivamente tres estructuras diferentes que llevaron a la obtención de los resultados que se describieron anteriormente, tomando la base de datos completa. Estos modelos fueron desarrollados considerando distintas arquitecturas de redes neuronales convolucionales unido a un perceptrón multicapa, con el fin de identificar la configuración que ofreciera el mejor rendimiento en la clasificación de los estadios de la DMAE. Básicamente, las primeras pruebas se llevaron a cabo siguiendo seis etapas en la primera fase del trabajo de grado:

1. Etapa 1: Carga de la base de datos.
2. Etapa 2: Preprocesamiento y procesamiento de las imágenes.
3. Etapa 3: Definición del modelo de CNN.
4. Etapa 4: Entrenamiento del modelo.
5. Etapa 5: Evaluación del modelo.
6. Etapa 6: Compilación de resultados.

A continuación, se presenta una muestra de la base de datos de imágenes obtenidas, en las que se diferencian 5 clases:

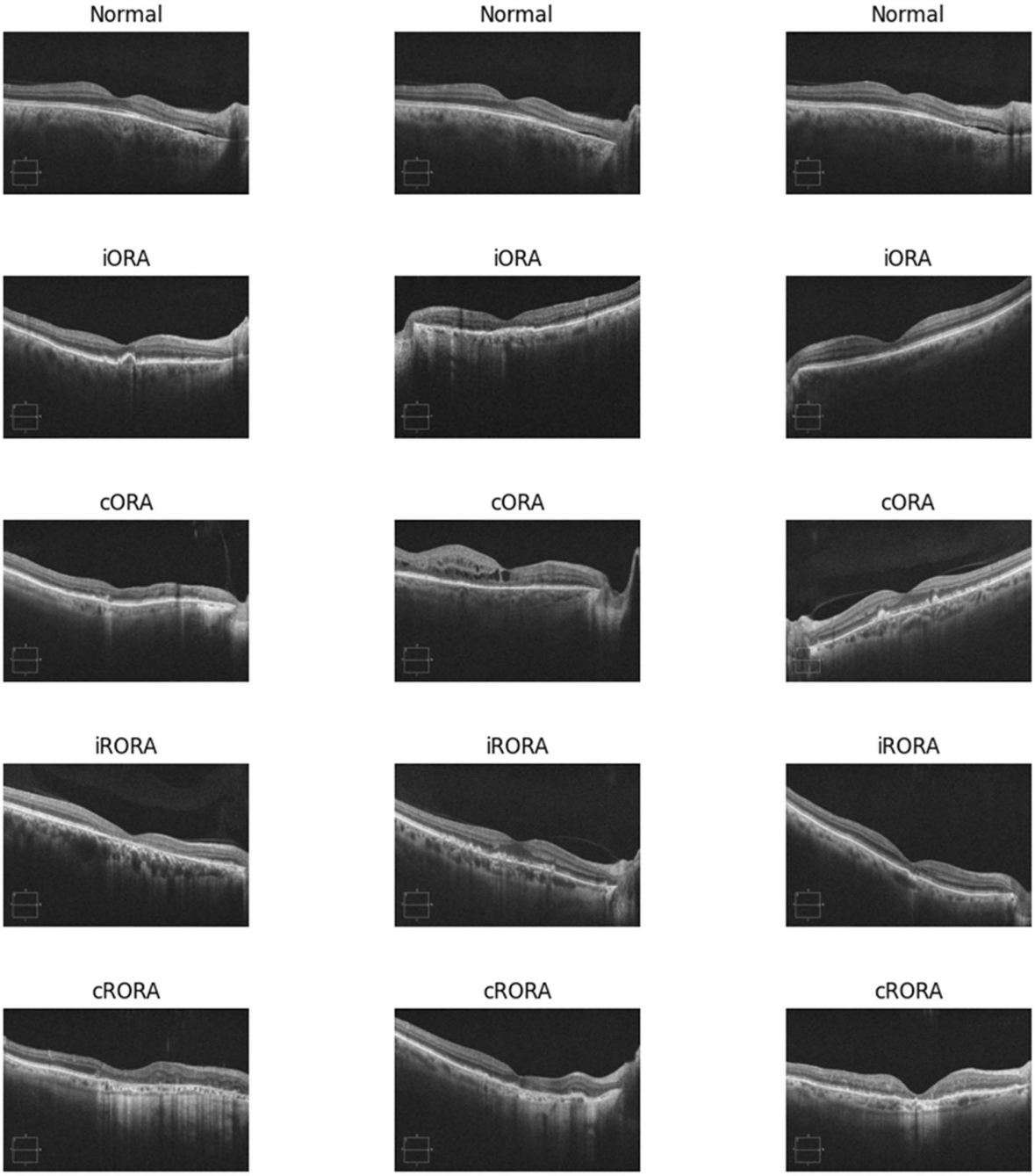


Figura No. 16 *– muestra de 3 imágenes por clase de la base de datos de estudio de la versión DMAE_A_Quiñones_Original_v4.ipynb*

Inicialmente, se realizaron pruebas con una CNN propia completa y con arquitecturas preentrenadas como ResNet50, EfficientNet y VGG16. Sin embargo, dado que los resultados no superaban el 68% de exactitud, se procedió a utilizar estas arquitecturas como extractores de características y usar algoritmos de aprendizaje automático clásico, tales como K-NN, SVM, Random Forest, entre otros. Es importante mencionar que las pruebas iniciales tenían una

duración considerable, aproximadamente entre 6 y 8 horas por modelo. Como se mencionó anteriormente, el modelo *'DMAE_A_Quiñones_Original_v4.ipynb*' logró la mejor exactitud hasta ese momento, alcanzando un 71%. Posteriormente, se experimentó con un proceso más eficiente, almacenando al modelo entrenado, y ejecutando los siguientes pasos:

1. Paso 1: Carga de la red convolucional entrenada en formato .keras (/content/drive/My Drive/DMAE/DMAE_A_Quiñones_Original_v4.keras).
2. Paso 2: Preparación de los datos para la extracción de características.
3. Paso 3: Extracción de las características de las imágenes y conversión de las listas en matrices NumPy.
4. Paso 4: División de los datos en conjuntos de entrenamiento y prueba.
5. Paso 5: Clasificación algoritmos de aprendizaje automático (K-NN, SVM, etc).
6. Paso 6: Evaluación del rendimiento del clasificador.

En esta nueva secuencia, se evaluó todo el modelo con las métricas previamente establecidas (exactitud, sensibilidad, especificidad, F1-Score y AUC-ROC), mejorando el costo computacional y reduciendo la duración de ejecución aproximada por clasificador a entre 30 minutos y 2 horas, ya que no era necesario ejecutar el entrenamiento de las redes convolucionales.

En la fase siguiente, se almacenaron las variables en un archivo comprimido .npz utilizando np.savez('DB_CNNpropia.npz', features=features, labels=labels). Este archivo sirvió como insumo para la última fase de experimentación, donde se trabajó con una estructura aún más simplificada, debido a que el archivo .npz contenía la base de datos completamente caracterizada:

1. Importación de librerías.
2. Carga de datos desde el Drive (datos = np.load('/content/drive/My Drive/DMAE/DB_CNNpropia.npz')).
3. División de datos y clasificación: Se aplicaron los pasos de división de los datos en entrenamiento y prueba, clasificación y evaluación del rendimiento del clasificador.
4. Evaluación del modelo mediante las métricas establecidas.

Esta estrategia permitió reducir aún más los tiempos de ejecución a entre 3 segundos a 20 minutos, dependiendo del algoritmo de aprendizaje automático. Además de la reducción de los tiempos, con los algoritmos básicos de aprendizaje automático se disminuyó el sobreajuste, y se incrementaron los índices de clasificación en el conjunto de validación, en comparación con el esquema inicial de CNN + Perceptrón.

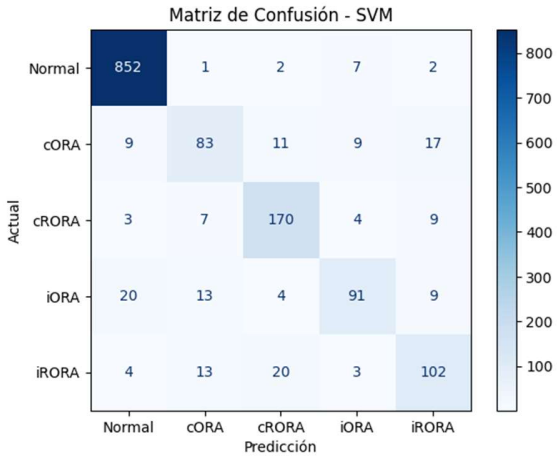
5. RESULTADOS Y DISCUSIÓN

5.1. EXTRACCIÓN DE CARACTERÍSTICAS CON CNN PROPIA PRE-ENTRENADA

5.1.1. CNN PROPIA + CLASIFICACIÓN CON SVM

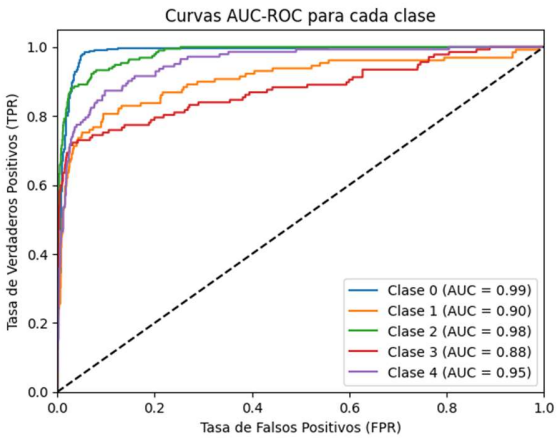
Esta etapa se llevó a cabo utilizando un clasificador de Máquina de Vectores de Soporte (SVM) con kernel radial (RBF). La SVM se entrenó con las características extraídas con la CNN propia. Posteriormente, se evaluó el rendimiento del clasificador utilizando varias métricas, como exactitud, matriz de confusión, AUC-ROC, precisión, recall, F1-Score, Balanced Accuracy y el coeficiente de Cohen's Kappa. La inclusión de estas métricas proporcionó una evaluación integral del modelo, permitiendo identificar tanto el rendimiento general como la capacidad de clasificación específica para cada clase de DMAE. Adicionalmente, se implementó una evaluación mediante curvas AUC-ROC para cada clase, análisis de sensibilidad y especificidad, así como el cálculo de log-loss, que mide el error del modelo al predecir probabilidades. Estas métricas permitieron obtener una visión más profunda del comportamiento del clasificador, destacando tanto sus fortalezas como posibles áreas de mejora. A pesar de la información que generaron la variedad de métricas exploradas, se estableció que, para efectos del alcance del proyecto se continuaría el análisis con accuracy, sensibilidad, especificidad, F1-score, y AUC-ROC. Esta combinación de CNN propia con SVM logró el mejor resultado promedio con una exactitud del 87.80% (Tabla No. 6), y una desviación estándar de 0.55.

Figura No. 17 *– Resultados de CNN Propia + Clasificación con SVM*



Reporte de Clasificación:	precision	recall	f1-score	support
Normal	0.96	0.99	0.97	864
cORA	0.71	0.64	0.67	129
cRORA	0.82	0.88	0.85	193
iORA	0.80	0.66	0.73	137
iRORA	0.73	0.72	0.73	142
accuracy			0.89	1465
macro avg	0.80	0.78	0.79	1465
weighted avg	0.88	0.89	0.88	1465

Ejemplo de una de las pruebas de la combinación CNN + SVM.



Métricas modelo SVM:

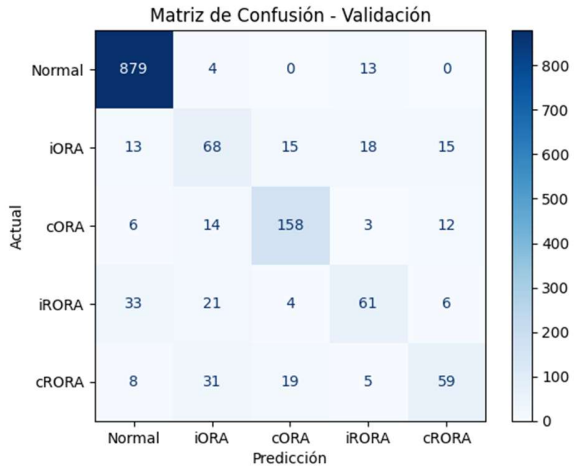
Exactitud: 88.60%
AUC-ROC General: 0.94
Log Loss: 0.0971
Cohen's Kappa: 0.81
Balanced Accuracy: 0.78
F1-Score Ponderado: 0.88
Sensibilidad General: 0.89
Especificidad Clase 1: 0.94
Especificidad Clase 2: 0.97
Especificidad Clase 3: 0.97
Especificidad Clase 4: 0.98
Especificidad Clase 5: 0.97
Especificidad General: 0.97

5.1.2. CNN PROPIA + CLASIFICACIÓN CON KNN-3

Se utilizan las características extraídas de la CNN propia con un clasificador K-Nearest Neighbors (K-NN) para identificación de estadios de la atrofia geográfica. El clasificador K-NN se ajusta con 3 vecinos (‘*n_neighbors=3*’), lo que significa que las predicciones se realizan basándose en la mayoría de los votos de los tres puntos más cercanos en el espacio de características. El uso de K-NN es adecuado en este contexto, ya que puede aprovechar las representaciones bien estructuradas de las características extraídas por la CNN, permitiendo una clasificación basada en la proximidad en un espacio de características de alta dimensionalidad.

La evaluación del modelo incluye la matriz de confusión, el reporte de clasificación y otras métricas de rendimiento. Los resultados muestran una exactitud global promedio del 82.69% (Tabla No. 6), con un AUC-ROC general de 0.89, lo que indica una buena capacidad del clasificador para diferenciar entre las cinco clases. Sin embargo, se observa variabilidad en el rendimiento por clase, siendo "*Normal*" la clase mejor predicha con un F1-score de 0.96, mientras que las clases con signos tempranos de atrofia como iORA e iRORA muestran menor precisión y recall, reflejando la dificultad del modelo para clasificar estas categorías menos diferenciadas. El análisis de la sensibilidad y especificidad muestra una sensibilidad general de 0.83 y una especificidad promedio de 0.95, que se traduce en una buena capacidad del modelo para identificar correctamente los casos positivos, aunque muestra un desempeño variable al diferenciar clases específicas.

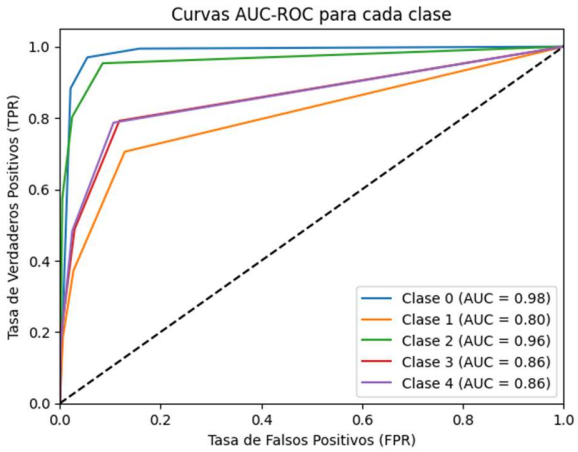
Figura No. 18 *– Resultados de CNN Propia + Clasificación con KNN-3*




```
Reporte de Clasificación:
              precision    recall  f1-score   support

      Normal       0.94      0.98      0.96       896
       iORA        0.49      0.53      0.51       129
       cORA        0.81      0.82      0.81       193
      iRORA        0.61      0.49      0.54       125
      cRORA        0.64      0.48      0.55       122

    accuracy                           0.84      1465
   macro avg       0.70      0.66      0.67      1465
weighted avg       0.83      0.84      0.83      1465
```



<u>Métricas modelo KNN-3</u>:

Exactitud del modelo K-NN: 83.62%
AUC-ROC General: 0.89
Log Loss: 5.9882
Cohen's Kappa: 0.71
Balanced Accuracy: 0.66
F1-Score Ponderado: 0.83
Sensibilidad General: 0.84
Especificidad Clase 1: 0.89
Especificidad Clase 2: 0.95
Especificidad Clase 3: 0.97
Especificidad Clase 4: 0.97
Especificidad Clase 5: 0.98
Especificidad General: 0.95

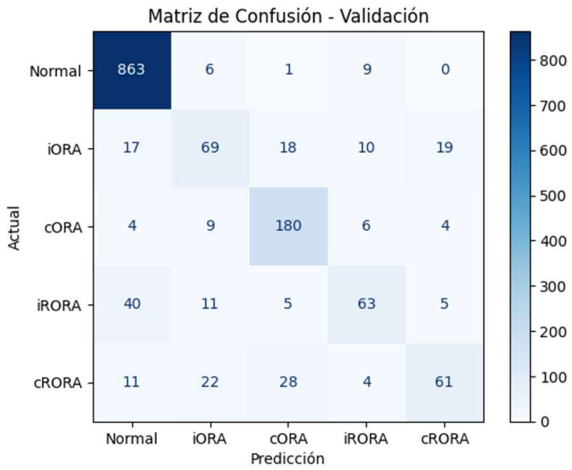
5.1.3. CNN PROPIA + CLASIFICACIÓN CON RANDOM FOREST

Se cargan las características extraídas previamente con la CNN propia pre-entrenada. El clasificador Random Forest, está configurado con 100 estimadores (‘*n_estimators=100*’). Random Forest es una técnica basada en el ensamble de múltiples árboles de decisión, lo que permite manejar mejor la varianza del modelo y mejorar su capacidad de generalización al combinar las predicciones de varios árboles.

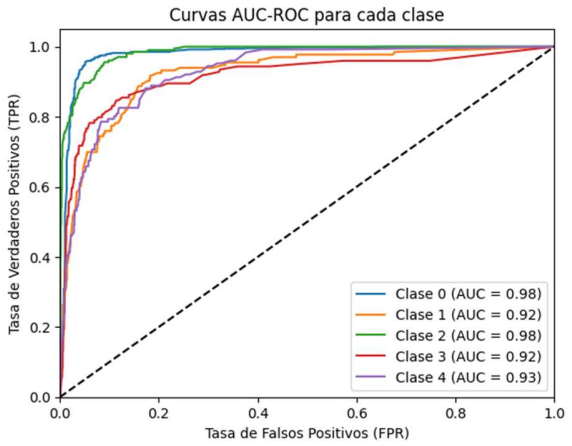
El modelo alcanza una exactitud del 84.57% (Tabla No. 6), con un AUC-ROC general de 0.95, lo que indica un excelente desempeño en la diferenciación entre las cinco clases estudiadas. La clase "*Normal*" muestra el mejor rendimiento, con un F1-score de 0.95 y una alta precisión, mientras que las clases relacionadas con estadios iniciales de atrofia (iORA e iRORA) presentan menor recall, reflejando una mayor dificultad en la identificación de estas categorías. El análisis de sensibilidad y especificidad muestra una sensibilidad general de 0.84 y una especificidad

promedio de 0.95, lo que resalta la alta capacidad del modelo para identificar correctamente los casos positivos y minimizar los falsos positivos. Sin embargo, se observan diferencias en el rendimiento por clase, siendo la especificidad particularmente alta en las clases de atrofia más avanzada (cORA y cRORA), lo que sugiere que el modelo maneja mejor estas categorías bien definidas en comparación con las etapas tempranas de la enfermedad.

Figura No. 19 *– Resultados de CNN Propia + Clasificación con Random Forest*



Reporte de Clasificación:	precision	recall	f1-score	support
Normal	0.92	0.98	0.95	879
iORA	0.59	0.52	0.55	133
cORA	0.78	0.89	0.83	203
iRORA	0.68	0.51	0.58	124
cRORA	0.69	0.48	0.57	126
accuracy			0.84	1465
macro avg	0.73	0.68	0.70	1465
weighted avg	0.83	0.84	0.83	1465



<u>Métricas modelo Random Forest</u>:

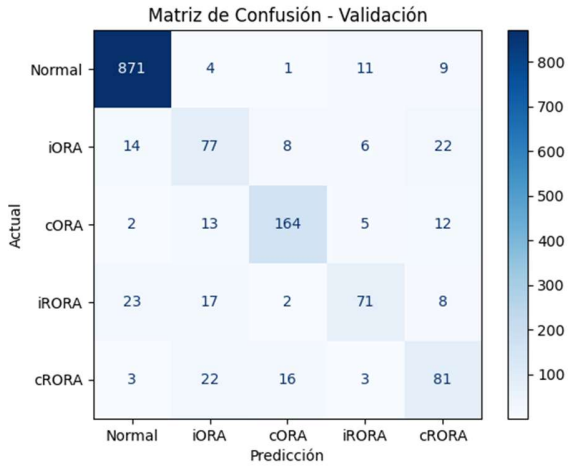
Exactitud del modelo Random Forest: 84.37%
AUC-ROC General: 0.95
F1-Score Ponderado: 0.83
Sensibilidad General: 0.84
Especificidad Clase 1: 0.88
Especificidad Clase 2: 0.96
Especificidad Clase 3: 0.96
Especificidad Clase 4: 0.98
Especificidad Clase 5: 0.98
Especificidad General: 0.95

5.1.4. CNN PROPIA + CLASIFICACIÓN CON XG-BOOST

La clasificación se realiza con XGBoost, configurado con parámetros estándar (‘*use_label_encoder=False*’, ’*eval_metric='logloss*'’). XGBoost utiliza el algoritmo de boosting, que combina múltiples modelos débiles para formar un modelo robusto y preciso, ajustando los errores de manera iterativa y mejorando la capacidad del clasificador para identificar patrones complejos. El modelo alcanza una exactitud general del 86.17% (Tabla No. 6) y un AUC-ROC de

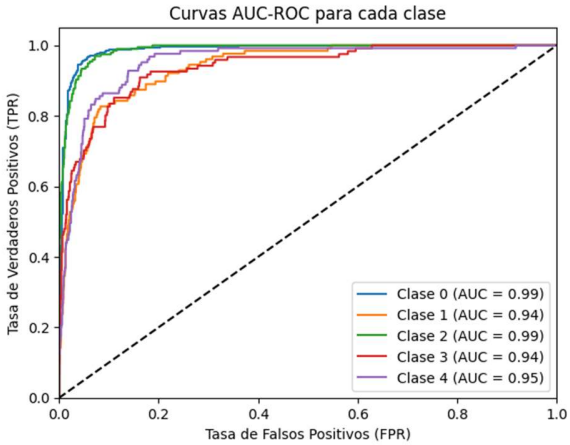
0.96, lo que indica un excelente desempeño en la diferenciación de las cinco clases mencionadas. La clase "*Normal*", como se podría esperar por el volumen de imágenes, mostró el mejor rendimiento, con una precisión y recall altos (0.95 y 0.97, respectivamente), mientras que las clases más complejas, como iORA y cRORA, presentaron menor precisión y recall, reflejando la dificultad del modelo para identificar estadios tempranos o menos diferenciados de la patología. El análisis de sensibilidad y especificidad muestra una sensibilidad general del 0.86 y una especificidad promedio del 0.96, lo que indica una alta capacidad del modelo para identificar correctamente tanto los casos positivos como los negativos. La especificidad es particularmente alta en las clases cORA y iRORA, sugiriendo que el modelo maneja mejor estas etapas avanzadas de atrofia. El F1-score ponderado de 0.86 refleja un desempeño equilibrado del clasificador en todas las categorías.

Figura No. 20 *– Resultados de CNN Propia + Clasificación con XG-Boost*



Reporte de Clasificación:

	precision	recall	f1-score	support
Normal	0.95	0.97	0.96	896
iORA	0.58	0.61	0.59	127
cORA	0.86	0.84	0.85	196
iRORA	0.74	0.59	0.65	121
cRORA	0.61	0.65	0.63	125
accuracy			0.86	1465
macro avg	0.75	0.73	0.74	1465
weighted avg	0.86	0.86	0.86	1465



<u>Métricas modelo XG-Boost</u>:

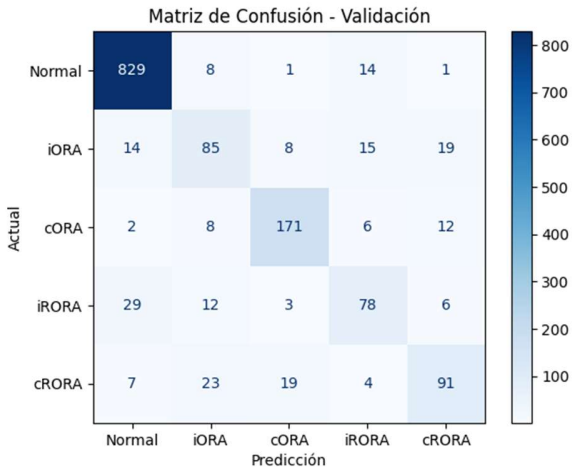
Exactitud del modelo XGBoost: 86.28%
AUC-ROC General: 0.96
F1-Score Ponderado: 0.86
Sensibilidad General: 0.86
Especificidad Clase 1: 0.93
Especificidad Clase 2: 0.96
Especificidad Clase 3: 0.98
Especificidad Clase 4: 0.98
Especificidad Clase 5: 0.96
Especificidad General: 0.96

5.1.5. CNN PROPIA + CLASIFICACIÓN CON GBM

En la fase de clasificación, se aplica GBM configurado con 100 estimadores, una tasa de aprendizaje de 0.1 y una profundidad máxima de 3 para los árboles. GBM es un método de ensamble que construye los modelos de forma secuencial, donde cada modelo ajusta los errores de su predecesor, mejorando iterativamente la precisión del clasificador. Esta técnica es especialmente efectiva para manejar datos complejos y detectar patrones no lineales en las características proporcionadas por la CNN.

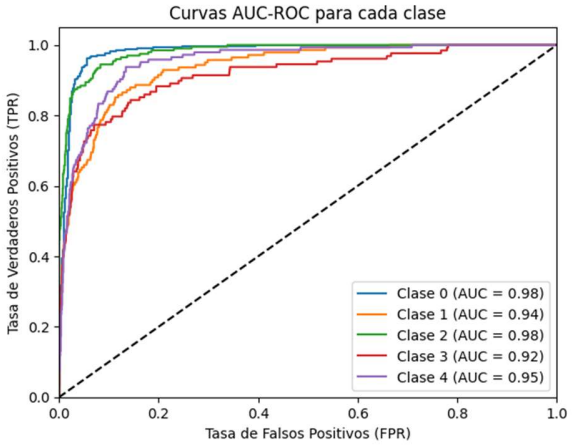
El modelo GBM alcanza una exactitud general del 85.60% (Tabla No. 6), con un AUC-ROC de 0.96, lo que indica un excelente rendimiento en la distinción de las cinco clases. La clase "*Normal*", al igual que en los modelos anteriores, muestra el mejor desempeño, con alta precisión y recall (0.94 y 0.97, respectivamente), mientras que las clases iORA e iRORA presentan menor rendimiento, reflejando la dificultad del modelo para clasificar correctamente estos estadios más complejos de la retinopatía en estudio. El análisis de sensibilidad y especificidad revela una sensibilidad general de 0.86 y una especificidad promedio de 0.96, indicando que el modelo tiene una alta capacidad para identificar tanto casos positivos como negativos. La alta especificidad observada en las clases cORA e iRORA sugiere que el modelo maneja mejor estas categorías avanzadas de atrofia. El F1-score ponderado de 0.85 refleja un buen balance en el desempeño general del clasificador, mientras que el macro promedio indica un rendimiento consistente a lo largo de todas las clases.

Figura No. 21 *– Resultados de CNN Propia + Clasificación con GBM*



Reporte de Clasificación:

	precision	recall	f1-score	support
Normal	0.94	0.97	0.96	853
iORA	0.62	0.60	0.61	141
cORA	0.85	0.86	0.85	199
iRORA	0.67	0.61	0.64	128
cRORA	0.71	0.63	0.67	144
accuracy			0.86	1465
macro avg	0.76	0.74	0.75	1465
weighted avg	0.85	0.86	0.85	1465



Métricas modelo GBM:

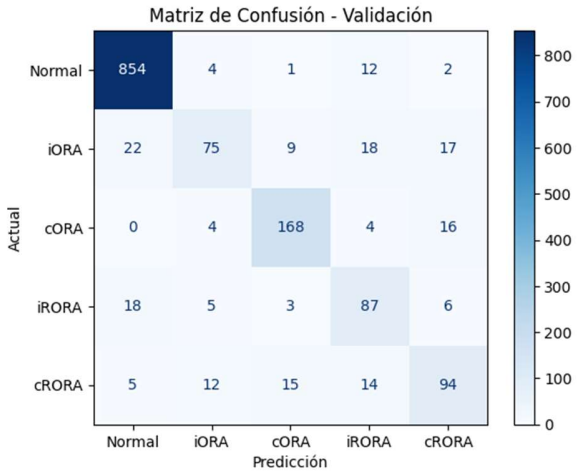
Exactitud del modelo Gradient Boosting Machines: 85.60%
AUC-ROC General: 0.95
F1-Score Ponderado: 0.85
Sensibilidad General: 0.86
Especificidad Clase 1: 0.92
Especificidad Clase 2: 0.96
Especificidad Clase 3: 0.98
Especificidad Clase 4: 0.97
Especificidad Clase 5: 0.97
Especificidad General: 0.96

5.1.6. CNN PROPIA + CLASIFICACIÓN CON MLP FEED FORWARD

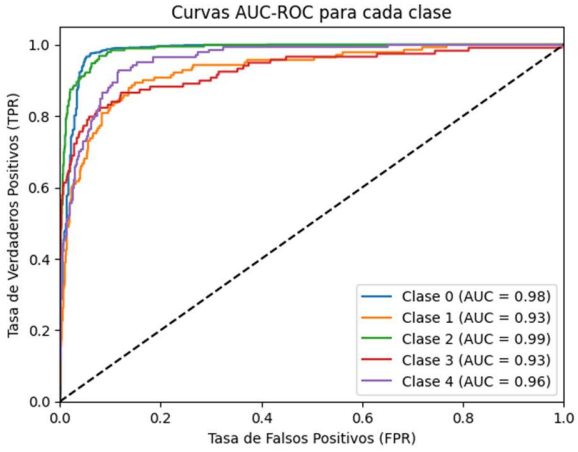
Esta red *fully-connected* incluye 4 capas densas (ocultas) de tamaño decreciente (1024, 512, 256 y 128 unidades), con activaciones ReLU para asegurar no linealidad y capas de Batch Normalization para estabilizar el proceso de entrenamiento. Además, se aplica Dropout al 50% para cada capa, mejorando la regularización y reduciendo el riesgo de sobreajuste. La última capa emplea una activación softmax para realizar la clasificación multiclase entre las cinco categorías estudiadas de la DMAE.

El entrenamiento se realiza con optimización Adam, usando una tasa de aprendizaje de 0.001, durante 30 épocas, con un tamaño de lote de 32. La evaluación del modelo se realiza en el conjunto de prueba, mostrando una exactitud del 87.25% (Tabla No. 6), lo que indica un buen desempeño del modelo. El reporte de clasificación muestra una alta precisión y recall para la clase "*Normal*" (clase con mayor volumen de datos) (0.95 y 0.98, respectivamente), mientras que las clases iORA e iRORA presentan menor rendimiento, reflejando la dificultad en la distinción de estas etapas intermedias de la DMAE, probablemente por el número limitado de imágenes en estas clases. El análisis adicional incluye métricas como el F1-Score ponderado (0.87) y AUC-ROC (0.96), sugiriendo un buen equilibrio entre precisión y recall, así como una alta capacidad de discriminación entre clases. La sensibilidad general del modelo es de 0.87, y la especificidad general es de 0.96, lo que indica una alta efectividad tanto en la detección de verdaderos positivos como en la correcta identificación de negativos. La especificidad por clase también es elevada, especialmente para las clases cORA e iRORA, con valores cercanos al 0.98.

Figura No. 22 *– Resultados de CNN Propia + Clasificación con MPL Feed Forward*



Reporte de Clasificación:	precision	recall	f1-score	support
Normal	0.95	0.98	0.96	873
iORA	0.75	0.53	0.62	141
cORA	0.86	0.88	0.87	192
iRORA	0.64	0.73	0.69	119
cRORA	0.70	0.67	0.68	140
accuracy			0.87	1465
macro avg	0.78	0.76	0.76	1465
weighted avg	0.87	0.87	0.87	1465



<u>Métricas modelo FeedFwrd</u>:

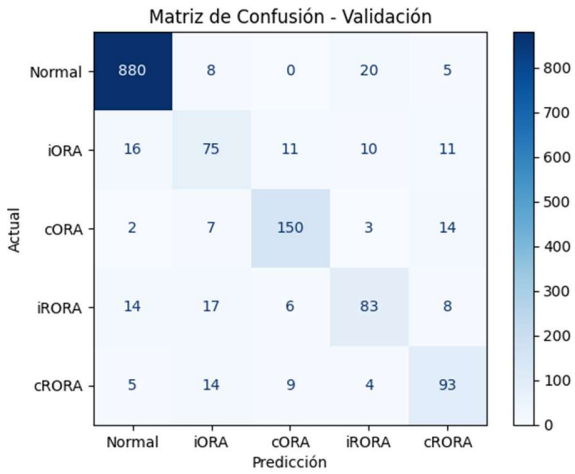
Exactitud del modelo Feedforward Neural Network: 87.24%
AUC-ROC General: 0.96
F1-Score Ponderado: 0.87
Sensibilidad General: 0.87
Especificidad Clase 1: 0.92
Especificidad Clase 2: 0.98
Especificidad Clase 3: 0.98
Especificidad Clase 4: 0.96
Especificidad Clase 5: 0.97
Especificidad General: 0.96

5.1.7. CNN PROPIA + CLASIFICACIÓN CON RNN

La arquitectura de la RNN consta de una capa SimpleRNN con 64 unidades y activación ReLU, seguida de una capa densa de 32 unidades. La capa recurrente se encarga de procesar la secuencia temporal de características, aprendiendo dependencias complejas en los datos extraídos. Además, se incluye una capa Dropout con una tasa de 0.4 para mejorar la regularización y prevenir el sobreajuste, junto con una capa de salida Softmax que realiza la clasificación multiclase. El modelo es compilado con el optimizador Adam y entrena durante 30 épocas con un tamaño de lote de 16, utilizando *sparse categorical crossentropy* como función de pérdida. El rendimiento del modelo se evaluó en el conjunto de prueba, mostrando una exactitud promedio del 87.04% (Tabla No. 6). El análisis de la matriz de confusión y el reporte de clasificación indica un rendimiento sólido para la clase "*Normal*" (precisión y recall de 0.96), reflejando una alta capacidad para detectar imágenes sin patología. Las clases intermedias como iORA e iRORA presentan menor desempeño, con recall de 0.61 y 0.65 respectivamente, lo que sugiere dificultad

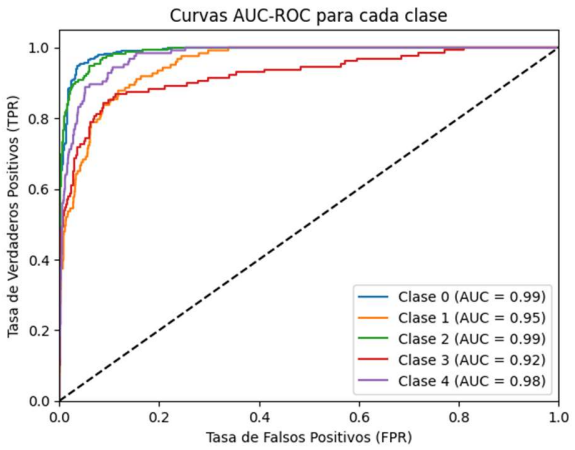
en la diferenciación de estas etapas en la progresión de la degeneración macular. El F1-Score ponderado es de 0.87, destacando un buen balance entre precisión y recall a nivel general. El AUC-ROC general es de 0.97, lo que indica una excelente capacidad discriminativa del modelo para distinguir entre las diferentes clases. La especificidad general alcanza 0.96, lo que sugiere un alto rendimiento en la identificación correcta de negativos. La sensibilidad general es de 0.87, evidenciando una buena capacidad para identificar verdaderos positivos.

Figura No. 23 *– Resultados de CNN Propia + Clasificación con RNN*



Reporte de Clasificación:

	precision	recall	f1-score	support
Normal	0.96	0.96	0.96	913
iORA	0.62	0.61	0.61	123
cORA	0.85	0.85	0.85	176
iRORA	0.69	0.65	0.67	128
cRORA	0.71	0.74	0.73	125
accuracy			0.87	1465
macro avg	0.77	0.76	0.76	1465
weighted avg	0.87	0.87	0.87	1465



<u>Métricas modelo RNN</u>:

Exactitud del modelo RNN: 87.44%
AUC-ROC General: 0.97
F1-Score Ponderado: 0.87
Sensibilidad General: 0.87
Especificidad Clase 1: 0.93
Especificidad Clase 2: 0.97
Especificidad Clase 3: 0.98
Especificidad Clase 4: 0.97
Especificidad Clase 5: 0.97
Especificidad General: 0.96

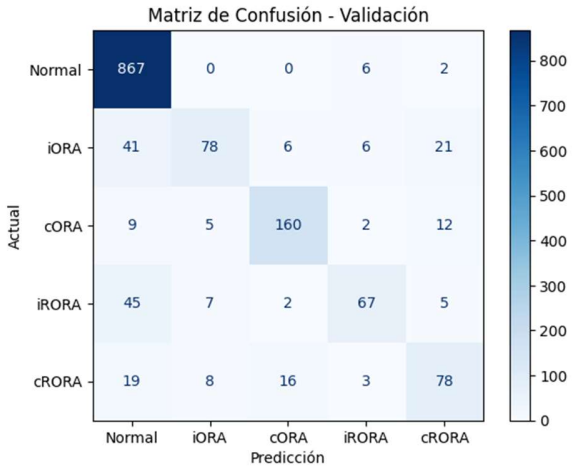
5.1.8. CNN PROPIA + CLASIFICACIÓN CON ENSEMBLEVOTING (SVM, FeedForward, RNN)

El ensamblaje de modelos se compone de tres clasificadores principales: un clasificador SVM con kernel RBF, una red Feedforward (FFNN) con dos capas ocultas (512 y 256 neuronas), y una Red Neuronal Recurrente (RNN) con capa SimpleRNN. El SVM y la FFNN son combinados mediante un Voting Classifier con votación suave (soft voting), mientras que las predicciones de la RNN se promedian con el Voting Classifier para obtener la predicción final. Esta estrategia busca aprovechar las fortalezas de cada clasificador, combinando la capacidad de decisión de la SVM, la

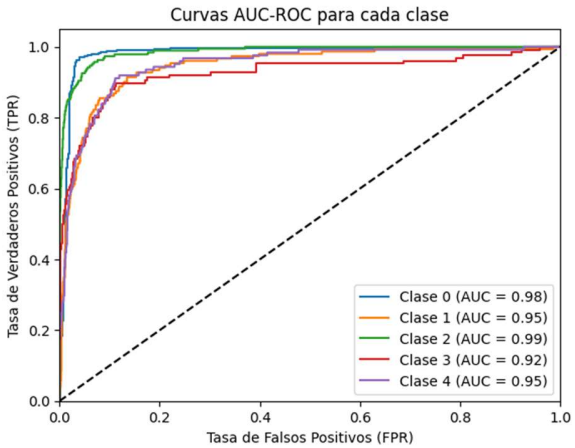
flexibilidad de la FFNN para aprender patrones no lineales, y la capacidad de la RNN para captar dependencias secuenciales en los datos.

La exactitud alcanzada fue del 86.37% (Tabla No. 6), con un AUC-ROC general de 0.96. El análisis de la matriz de confusión muestra un excelente rendimiento para la clase "*Normal*", con alta precisión y recall (0.88 y 0.99 respectivamente), indicando una fuerte capacidad para identificar imágenes sin patología. Sin embargo, las clases más complejas, como iORA e iRORA, presentan menor recall (0.51 y 0.53 respectivamente), lo que sugiere dificultades en la distinción de estas etapas intermedias. El F1-Score ponderado es de 0.85, reflejando un balance adecuado entre precisión y recall a nivel general. La especificidad general es de 0.95, destacando un rendimiento robusto en la identificación de negativos, con valores particularmente altos para las clases cORA e iRORA. La sensibilidad general es de 0.85, evidenciando una buena capacidad para identificar verdaderos positivos.

Figura No. 24 *– Resultados de CNN Propia + Clasificación con EnsembleVoting (SVM, FeedForward, RNN)*



Reporte de Clasificación:	precision	recall	f1-score	support
Normal	0.88	0.99	0.93	875
iORA	0.80	0.51	0.62	152
cORA	0.87	0.85	0.86	188
iRORA	0.80	0.53	0.64	126
cRORA	0.66	0.63	0.64	124
accuracy			0.85	1465
macro avg	0.80	0.70	0.74	1465
weighted avg	0.85	0.85	0.84	1465



<u>Métricas modelo EnsembleVoting</u>:

Exactitud del modelo EnsembleVoting: 85.32%
AUC-ROC General: 0.96
F1-Score Ponderado: 0.84
Sensibilidad General: 0.85
Especificidad Clase 1: 0.81
Especificidad Clase 2: 0.98
Especificidad Clase 3: 0.98
Especificidad Clase 4: 0.99
Especificidad Clase 5: 0.97
Especificidad General: 0.95

CNN_Propia: Extractor de características

MODELO	K-NN (3)				
Repetición	Accuracy	Sensibilidad	Especificidad	F1-Score	AUC-ROC
1	83,62	0,84	0,95	0,83	0,89
2	83,14	0,83	0,95	0,82	0,90
3	80,41	0,80	0,94	0,80	0,88
4	82,66	0,83	0,95	0,82	0,89
5	83,62	0,84	0,95	0,83	0,89
Promedio	**82,69**	0,83	0,95	0,82	0,89
STD	1,34	0,02	0,00	0,01	0,01

MODELO	SVM				
Repetición	Accuracy	Sensibilidad	Especificidad	F1-Score	AUC-ROC
1	87,65	0,88	0,96	0,87	0,94
2	87,10	0,87	0,96	0,87	0,94
3	87,65	0,88	0,96	0,88	0,93
4	88,60	0,89	0,97	0,89	0,94
5	87,99	0,88	0,97	0,88	0,93
Promedio	**87,80**	0,88	0,96	0,88	0,94
STD	0,55	0,01	0,01	0,01	0,01

MODELO	Random Forest				
Repetición	Accuracy	Sensibilidad	Especificidad	F1-Score	AUC-ROC
1	84,98	0,85	0,96	0,84	0,95
2	84,03	0,84	0,95	0,83	0,95
3	84,23	0,84	0,95	0,83	0,95
4	85,26	0,85	0,95	0,84	0,95
5	84,37	0,84	0,95	0,83	0,95
Promedio	**84,57**	0,84	0,95	0,83	0,95
STD	0,52	0,01	0,00	0,01	0,00

MODELO	XG-Boost				
Repetición	Accuracy	Sensibilidad	Especificidad	F1-Score	AUC-ROC
1	87,37	0,87	0,96	0,87	0,96
2	86,21	0,86	0,96	0,86	0,96
3	85,05	0,85	0,96	0,85	0,95
4	85,94	0,86	0,96	0,86	0,95
5	86,28	0,86	0,96	0,86	0,96
Promedio	**86,17**	0,86	0,96	0,86	0,96
STD	0,83	0,01	0,00	0,01	0,01

MODELO	GBM				
Repetición	Accuracy	Sensibilidad	Especificidad	F1-Score	AUC-ROC
1	86,48	0,86	0,96	0,86	0,96
2	85,53	0,86	0,96	0,85	0,96
3	85,94	0,86	0,96	0,86	0,96
4	84,44	0,84	0,96	0,84	0,95
5	85,60	0,86	0,96	0,85	0,95
Promedio	**85,60**	0,86	0,96	0,85	0,96
STD	0,75	0,01	0,00	0,01	0,01

MODELO	Feed Forward				
Repetición	Accuracy	Sensibilidad	Especificidad	F1-Score	AUC-ROC
1	86,35	0,86	0,97	0,86	0,96
2	87,37	0,87	0,96	0,87	0,96
3	87,58	0,88	0,96	0,87	0,95
4	87,71	0,88	0,97	0,88	0,96
5	87,24	0,87	0,96	0,87	0,96
Promedio	**87,25**	0,87	0,96	0,87	0,96
STD	0,54	0,01	0,01	0,01	0,00

MODELO	RNN				
Repetición	Accuracy	Sensibilidad	Especificidad	F1-Score	AUC-ROC
1	86,35	0,86	0,96	0,86	0,96
2	86,48	0,86	0,96	0,86	0,96
3	87,17	0,87	0,96	0,87	0,96
4	87,78	0,88	0,96	0,88	0,96
5	87,44	0,87	0,96	0,87	0,97
Promedio	**87,04**	0,87	0,96	0,87	0,96
STD	0,62	0,01	0,00	0,01	0,00

MODELO	EnsembleVoting (SVM, FeedForward, RNN)				
Repetición	Accuracy	Sensibilidad	Especificidad	F1-Score	AUC-ROC
1	85,12	0,85	0,95	0,84	0,96
2	87,17	0,87	0,96	0,86	0,96
3	85,32	0,85	0,95	0,84	0,96
4	86,96	0,87	0,96	0,86	0,96
5	87,30	0,87	0,95	0,87	0,97
Promedio	**86,37**	0,86	0,95	0,85	0,96
STD	1,06	0,01	0,01	0,01	0,00

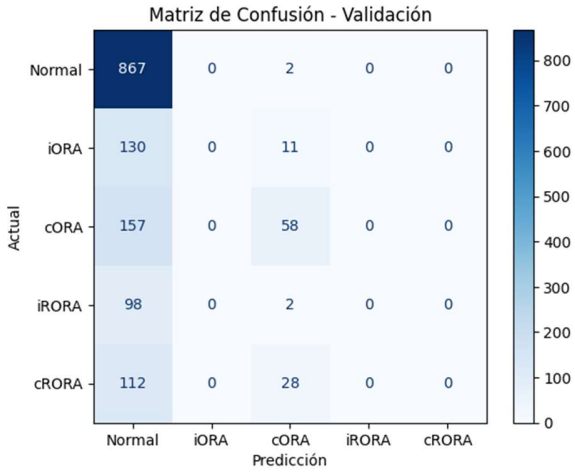
Tabla No. 6 *–* ***Pruebas con CNN Propia como extractor de características****. Se realizaron 5 repeticiones por clasificador validando promedios y desviaciones estándar.*

5.2. EXTRACCIÓN DE CARACTERÍTICAS CON ARQUITECTURAS PREENTRENADAS: VGG16

5.2.1. VGG16 + CLASIFICACIÓN CON SVM

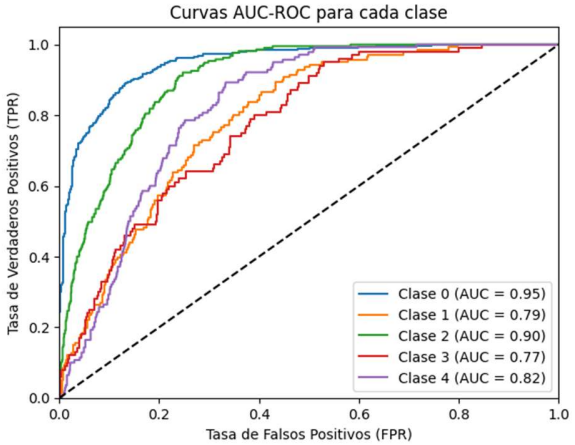
En los resultados, el modelo alcanzó una exactitud del 64.53% (Tabla No. 7), indicando un desempeño bajo. Aunque la clase "*Normal*" muestra buenos resultados con un F1-score de 0.78 en el reporte de clasificación y una precisión del 64%, las demás clases (iORA, cORA, iRORA y cRORA) presentan un desempeño deficiente, con F1-scores muy bajos, especialmente para iORA, iRORA y cRORA, que tienen un F1-score de 0.00. Este rendimiento indica que el modelo tiene dificultades significativas para clasificar las etapas de DMAE más avanzadas y posiblemente confunde estas clases debido a la similitud en las características visuales o al desbalance de datos. El AUC-ROC general alcanzó 0.85, lo que sugiere que el modelo tiene cierta capacidad discriminativa, aunque esta no se refleja claramente en la exactitud y el F1-score. La especificidad es alta para las clases iORA, iRORA y cRORA (1.00).

Figura No. 25 *– Resultados de VGG16 + Clasificación con SVM*



Reporte de Clasificación:

	precision	recall	f1-score	support
Normal	0.64	1.00	0.78	869
iORA	0.00	0.00	0.00	141
cORA	0.57	0.27	0.37	215
iRORA	0.00	0.00	0.00	100
cRORA	0.00	0.00	0.00	140
accuracy			0.63	1465
macro avg	0.24	0.25	0.23	1465
weighted avg	0.46	0.63	0.51	1465



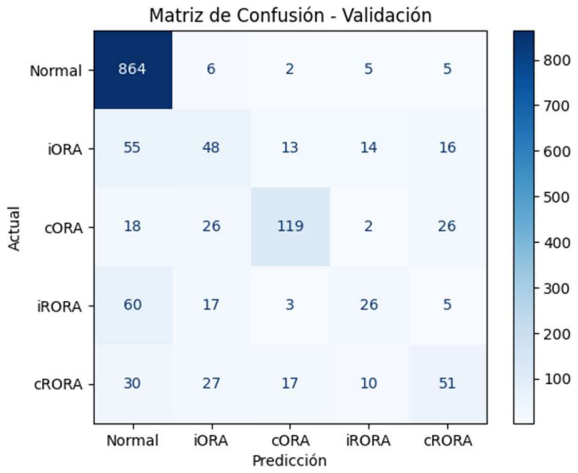
<u>Métricas modelo SVM</u>:

Exactitud del modelo SVM: 63.14%
AUC-ROC General: 0.85
F1-Score Ponderado: 0.51
Sensibilidad General: 0.63
Especificidad Clase 1: 0.17
Especificidad Clase 2: 1.00
Especificidad Clase 3: 0.97
Especificidad Clase 4: 1.00
Especificidad Clase 5: 1.00
Especificidad General: 0.83

5.2.2. VGG16 PRE-ENTRENADA + CLASIFICACIÓN CON KNN-3

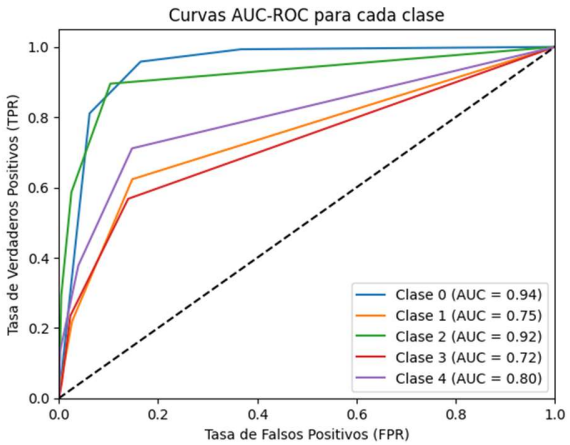
Los resultados muestran una exactitud promedio del 75.96% (Tabla No. 7), lo que indica un rendimiento moderado del clasificador K-NN. La clase "*Normal*" obtuvo los mejores resultados, con un F1-score de 0.91 y alta precisión (0.84). Sin embargo, las clases iORA, iRORA y cRORA presentan mayores dificultades, con un F1-score de 0.36, 0.31 y 0.43 respectivamente, reflejando problemas en la clasificación de estas etapas intermedias de la degeneración macular. Esto podría estar relacionado con la similitud visual entre estas clases y la posible insuficiencia de muestras representativas en el conjunto de datos. El AUC-ROC general alcanzó 0.83, lo que sugiere una capacidad discriminativa razonable del modelo, aunque con espacio para mejoras. La especificidad general fue alta (0.91), especialmente para las clases más avanzadas (cORA e iRORA). La sensibilidad general fue de 0.76, lo que muestra una moderada capacidad para detectar verdaderos positivos, con dificultades específicas en las clases menos representadas

Figura No. 26 *– Resultados de VGG16 + Clasificación con KNN-3*



Reporte de Clasificación:

	precision	recall	f1-score	support
Normal	0.84	0.98	0.91	882
iORA	0.39	0.33	0.36	146
cORA	0.77	0.62	0.69	191
iRORA	0.46	0.23	0.31	111
cRORA	0.50	0.38	0.43	135
accuracy			0.76	1465
macro avg	0.59	0.51	0.54	1465
weighted avg	0.73	0.76	0.73	1465



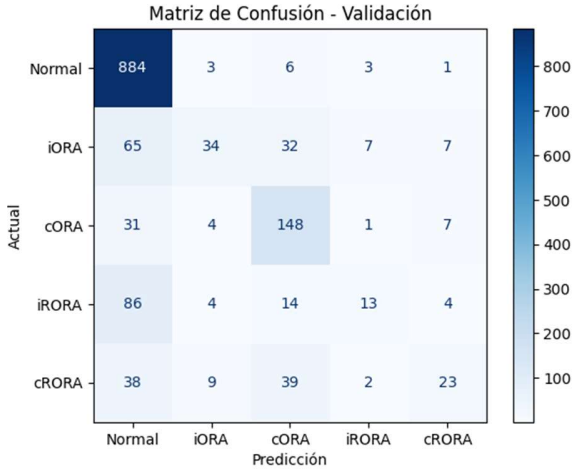
<u>Métricas modelo KNN-3</u>:

Exactitud del modelo K-NN: 75.63%
AUC-ROC General: 0.83
F1-Score Ponderado: 0.73
Sensibilidad General: 0.76
Especificidad Clase 1: 0.72
Especificidad Clase 2: 0.94
Especificidad Clase 3: 0.97
Especificidad Clase 4: 0.98
Especificidad Clase 5: 0.96
Especificidad General: 0.91

5.2.3. VGG16 PRE-ENTRENADA + CLASIFICACIÓN CON RANDOM FOREST

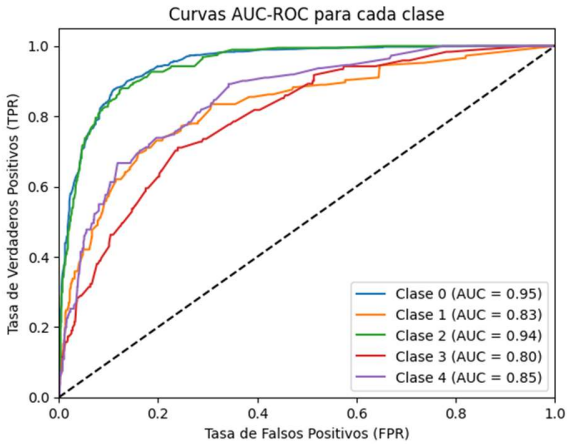
Los resultados muestran una exactitud del 74.37% (Tabla No. 7), lo que indica un desempeño moderado. La clase "*Normal*" tiene un rendimiento destacado con un F1-score de 0.88 (reporte de clasificación) y una alta precisión del 80%, mientras que las demás clases, especialmente iRORA y cRORA, presentan un rendimiento bajo, con F1-scores de 0.18 y 0.30, respectivamente. Esto sugiere que el modelo tiene dificultades para distinguir entre las etapas avanzadas de la patología estudiada, posiblemente debido a la similitud en las características visuales de estas clases y al desbalance en la distribución de datos. El AUC-ROC general alcanzó 0.87. La especificidad es alta para las clases más complejas (iORA, iRORA y cRORA), alcanzando valores de hasta 0.99. Sin embargo, la baja sensibilidad para estas clases muestra una limitación significativa en la detección de verdaderos positivos, especialmente en las etapas más avanzadas de DMAE.

Figura No. 27 *– Resultados de VGG16 + Clasificación con Random Forest*



Reporte de Clasificación:

	precision	recall	f1-score	support
Normal	0.80	0.99	0.88	897
iORA	0.63	0.23	0.34	145
cORA	0.62	0.77	0.69	191
iRORA	0.50	0.11	0.18	121
cRORA	0.55	0.21	0.30	111
accuracy			0.75	1465
macro avg	0.62	0.46	0.48	1465
weighted avg	0.72	0.75	0.70	1465



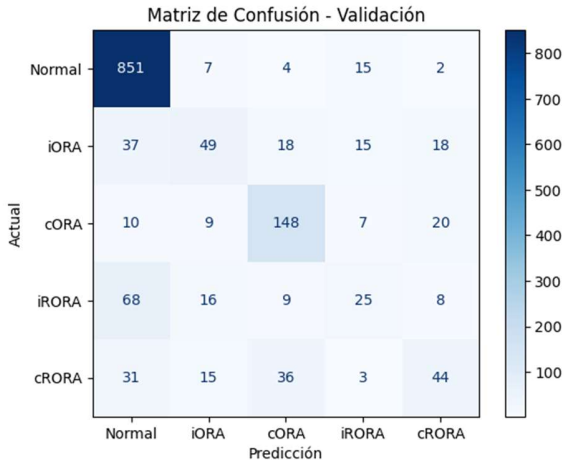
<u>Métricas modelo Random Forest</u>:

Exactitud del modelo Random Forest: 75.22%
AUC-ROC General: 0.87
F1-Score Ponderado: 0.70
Sensibilidad General: 0.75
Especificidad Clase 1: 0.61
Especificidad Clase 2: 0.98
Especificidad Clase 3: 0.93
Especificidad Clase 4: 0.99
Especificidad Clase 5: 0.99
Especificidad General: 0.90

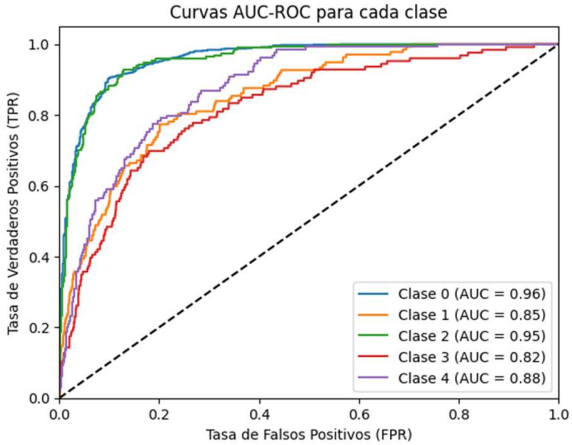
5.2.4. VGG16 PRE-ENTRENADA + CLASIFICACIÓN CON XG-BOOST

Los resultados muestran una exactitud del 76.48% en promedio (Tabla No. 7), lo que indica un rendimiento moderado. La clase "*Normal*" obtiene el mejor desempeño con un F1-Score de 0.91, reflejando tanto alta precisión como alta sensibilidad. Sin embargo, el rendimiento disminuye notablemente en las clases avanzadas de DMAE, como iRORA y cRORA, que muestran F1-Scores bajos (0.26 y 0.40, respectivamente). Esto apunta a que el modelo tiene dificultades para diferenciar entre estas etapas, posiblemente debido a similitudes visuales en las características extraídas y al desbalance de datos. El AUC-ROC general alcanza un valor de 0.89, indicando una capacidad aceptable para discriminar entre clases, aunque la discriminación no es uniforme en todas las categorías. La especificidad general es alta (0.92), lo que indica que el modelo identifica bien los casos negativos, pero la baja sensibilidad para las clases avanzadas evidencia problemas en la detección de verdaderos positivos.

Figura No. 28 *– Resultados de VGG16 + Clasificación con XG-Boost*



Reporte de Clasificación:	precision	recall	f1-score	support
Normal	0.85	0.97	0.91	879
iORA	0.51	0.36	0.42	137
cORA	0.69	0.76	0.72	194
iRORA	0.38	0.20	0.26	126
cRORA	0.48	0.34	0.40	129
accuracy			0.76	1465
macro avg	0.58	0.53	0.54	1465
weighted avg	0.73	0.76	0.74	1465



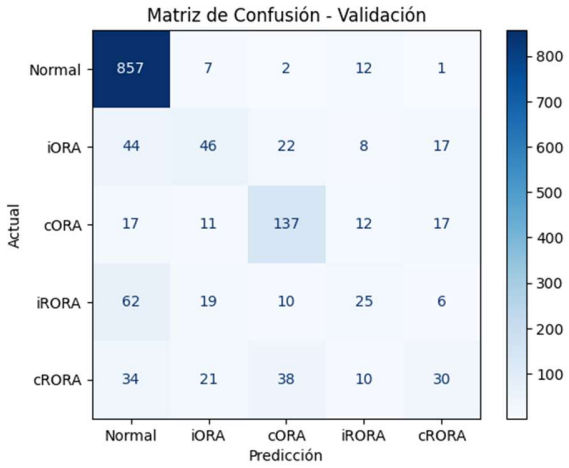
<u>Métricas modelo XG-Boost</u>:

Exactitud del modelo XGBoost: 76.25%
AUC-ROC General: 0.89
F1-Score Ponderado: 0.74
Sensibilidad General: 0.76
Especificidad Clase 1: 0.75
Especificidad Clase 2: 0.96
Especificidad Clase 3: 0.95
Especificidad Clase 4: 0.97
Especificidad Clase 5: 0.96
Especificidad General: 0.92

5.2.5. VGG16 PRE-ENTRENADA + CLASIFICACIÓN CON GBM

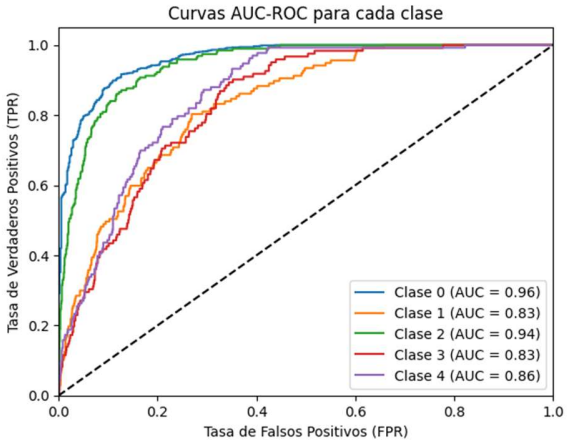
La matriz de confusión revela que el modelo tiene un desempeño adecuado en la clase "*Normal*", con una alta precisión (0.85) y un F1-Score de 0.91, indicando una correcta identificación de los casos sin retinopatía. Sin embargo, el rendimiento disminuye significativamente en las clases patológicas avanzadas (iORA, cRORA), reflejado en F1-Scores bajos (0.26 y 0.29, respectivamente). Esto sugiere que el modelo tiene dificultades para diferenciar entre etapas avanzadas de la DMAE, posiblemente debido a características visuales similares y al desbalance en el conjunto de datos. El AUC-ROC general es de 0.89, lo que indica una buena capacidad discriminativa del modelo, aunque no es uniforme para todas las clases. Las especificidades son altas para las clases avanzadas, lo que significa que el modelo logra identificar bien los casos negativos. No obstante, la baja sensibilidad para las mismas clases refleja problemas en la detección de verdaderos positivos, afectando así la capacidad del modelo para generalizar a nuevos datos.

Figura No. 29 *– Resultados de VGG16 + Clasificación con GBM*



Reporte de Clasificación:

	precision	recall	f1-score	support
Normal	0.85	0.97	0.91	879
iORA	0.44	0.34	0.38	137
cORA	0.66	0.71	0.68	194
iRORA	0.37	0.20	0.26	122
cRORA	0.42	0.23	0.29	133
accuracy			0.75	1465
macro avg	0.55	0.49	0.51	1465
weighted avg	0.70	0.75	0.72	1465



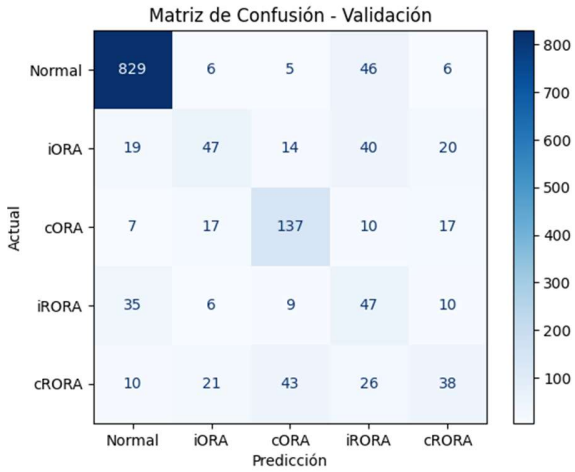
<u>Métricas modelo GBM</u>:

Exactitud del modelo GBM: 74.74%
AUC-ROC General: 0.89
F1-Score Ponderado: 0.72
Sensibilidad General: 0.75
Especificidad Clase 1: 0.73
Especificidad Clase 2: 0.96
Especificidad Clase 3: 0.94
Especificidad Clase 4: 0.97
Especificidad Clase 5: 0.97
Especificidad General: 0.91

5.2.6. VGG16 PRE-ENTRENADA + CLASIFICACIÓN CON MLP FEED FORWARD

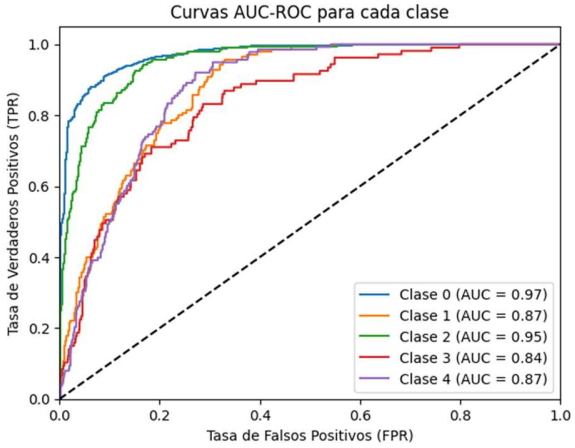
En cuanto a los resultados, el modelo presenta una exactitud de 74.61% (Tabla No. 7), con un AUC-ROC general de 0.90, lo que indica una buena capacidad de discriminación. La clase "*Normal*" muestra una alta precisión (0.92) y F1-Score (0.93), lo que refleja una correcta clasificación de esta categoría. Sin embargo, el rendimiento disminuye considerablemente en las clases patológicas, especialmente en "iRORA" y "cRORA", con F1-Scores de 0.34 y 0.33, respectivamente. Esto sugiere dificultades del modelo para identificar estas categorías, probablemente debido a la similitud visual entre las etapas avanzadas de DMAE y el desbalance de clases. La matriz de confusión revela que el modelo confunde frecuentemente "iORA" y "cORA", lo que podría ser abordado mediante técnicas adicionales de ajuste del modelo o aumento de datos. La especificidad es alta en general, alcanzando un 0.93, lo que indica que el modelo logra evitar falsos positivos de manera efectiva. Sin embargo, la sensibilidad general del 0.75 indica que todavía hay margen para mejorar en la identificación de verdaderos positivos, particularmente en las clases menos representadas.

Figura No. 30 *– Resultados de VGG16 + Clasificación con MPL Feed Forward*



Reporte de Clasificación:

	precision	recall	f1-score	support
Normal	0.92	0.93	0.93	892
iORA	0.48	0.34	0.40	140
cORA	0.66	0.73	0.69	188
iRORA	0.28	0.44	0.34	107
cRORA	0.42	0.28	0.33	138
accuracy			0.75	1465
macro avg	0.55	0.54	0.54	1465
weighted avg	0.75	0.75	0.75	1465



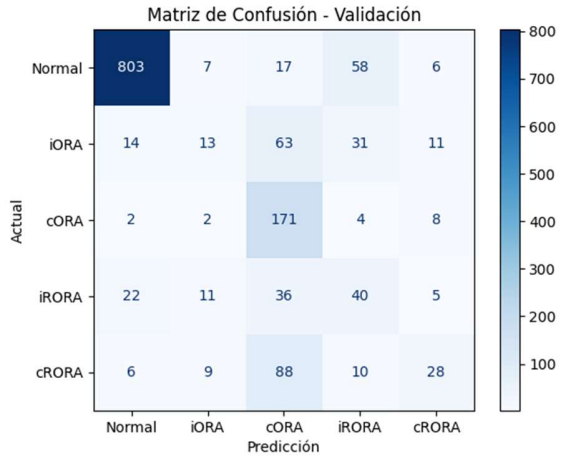
<u>Métricas modelo FeedFwrd</u>:

Exactitud del modelo Feed Forward: 74.95%
AUC-ROC General: 0.90
F1-Score Ponderado: 0.75
Sensibilidad General: 0.75
Especificidad Clase 1: 0.88
Especificidad Clase 2: 0.96
Especificidad Clase 3: 0.94
Especificidad Clase 4: 0.91
Especificidad Clase 5: 0.96
Especificidad General: 0.93

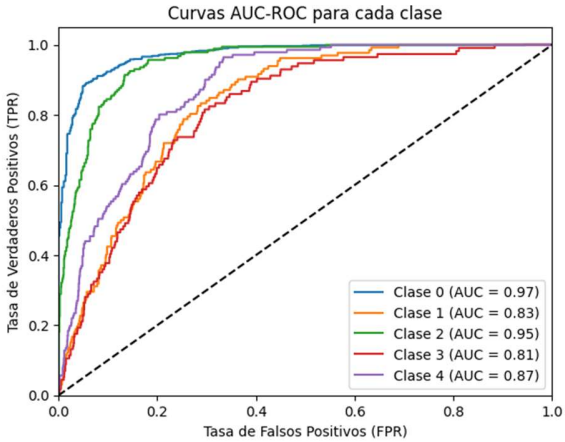
5.2.7. VGG16 PRE-ENTRENADA Y CLASIFICACIÓN CON RNN

El modelo RNN logró una exactitud general del 73.38% en promedio (Tabla No. 7) y un AUC-ROC promedio de 0.89, lo que indica una capacidad moderada para discriminar entre clases. La clase "*Normal*" mostró un alto desempeño con un F1-score de 0.92, mientras que el rendimiento en las clases patológicas fue considerablemente más bajo. Por ejemplo, "iORA" y "iRORA" presentaron F1-scores de 0.15 y 0.31, respectivamente, lo que sugiere una dificultad para identificar estas categorías debido a su similitud visual o a un desbalance en el conjunto de datos. El análisis de la matriz de confusión revela que el modelo tiene problemas para distinguir entre clases patológicas, especialmente "cORA" y "iORA", lo que podría deberse a la variabilidad en las presentaciones clínicas y la similitud en las características morfológicas. La especificidad general fue alta (0.93), lo que indica que el modelo maneja bien los falsos positivos. Sin embargo, la sensibilidad general fue moderada (0.72), destacando la necesidad de mejorar la identificación de verdaderos positivos.

Figura No. 31 *– Resultados de VGG16 + Clasificación con RNN*



Reporte de Clasificación:	precision	recall	f1-score	support
Normal	0.95	0.90	0.92	891
iORA	0.31	0.10	0.15	132
cORA	0.46	0.91	0.61	187
iRORA	0.28	0.35	0.31	114
cRORA	0.48	0.20	0.28	141
accuracy			0.72	1465
macro avg	0.50	0.49	0.45	1465
weighted avg	0.73	0.72	0.70	1465



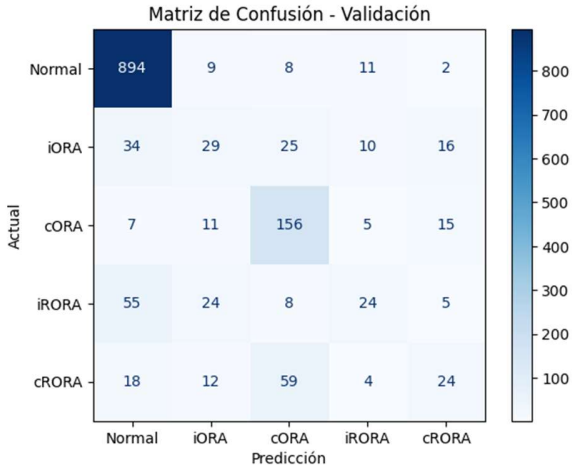
<u>Métricas modelo RNN</u>:

Exactitud del modelo RNN: 72.01%
AUC-ROC General: 0.89
F1-Score Ponderado: 0.70
Sensibilidad General: 0.72
Especificidad Clase 1: 0.92
Especificidad Clase 2: 0.98
Especificidad Clase 3: 0.84
Especificidad Clase 4: 0.92
Especificidad Clase 5: 0.98
Especificidad General: 0.93

5.2.8. VGG16 PRE-ENTRENADA + CLASIFICACIÓN CON ENSEMBLEVOTING (SVM, FeedForward, RNN)

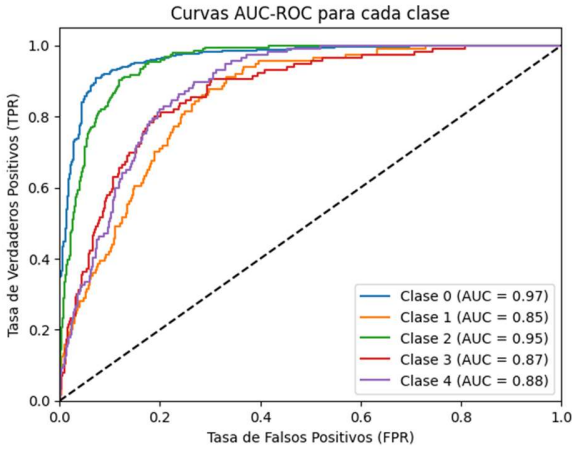
La exactitud general del ensamblaje fue del 74.95% en promedio (Tabla No. 7), con un AUC-ROC promedio de 0.90, indicando un buen rendimiento en la mayoría de las clases. El modelo mostró un desempeño destacado en la clase "Normal" con un F1-score de 0.93, reflejando una alta precisión y recall. Sin embargo, las clases patológicas, como "iORA" e "iRORA", presentaron dificultades, con F1-scores bajos (0.29 y 0.28, respectivamente). Esto sugiere que el modelo tiene problemas para discriminar estas categorías debido a su similitud visual o posible desbalance en los datos. La matriz de confusión muestra que la clase "Normal" es la mejor identificada, mientras que las clases patológicas tienden a confundirse entre sí. La especificidad general es alta (0.92), lo que indica una buena capacidad para evitar falsos positivos, especialmente en las clases patológicas. No obstante, la sensibilidad general fue moderada (0.77), lo que destaca la necesidad de mejorar la detección de verdaderos positivos, particularmente en las clases menos representadas.

Figura No. 32 *– Resultados de VGG16 + Clasificación con EnsembleVoting (SVM, FeedForward, RNN)*



Reporte de Clasificación:

	precision	recall	f1-score	support
Normal	0.89	0.97	0.93	924
iORA	0.34	0.25	0.29	114
cORA	0.61	0.80	0.69	194
iRORA	0.44	0.21	0.28	116
cRORA	0.39	0.21	0.27	117
accuracy			0.77	1465
macro avg	0.53	0.49	0.49	1465
weighted avg	0.73	0.77	0.74	1465



<u>Métricas modelo EnsembleVoting</u>:

Exactitud del modelo Ensemble Voting: 76.93%
AUC-ROC General: 0.90
F1-Score Ponderado: 0.74
Sensibilidad General: 0.77
Especificidad Clase 1: 0.79
Especificidad Clase 2: 0.96
Especificidad Clase 3: 0.92
Especificidad Clase 4: 0.98
Especificidad Clase 5: 0.97
Especificidad General: 0.92

VGG16: Extractor de características

MODELO	K-NN (3)				
Repetición	Accuracy	Sensibilidad	Especificidad	F1-Score	AUC-ROC
1	75,63	0,76	0,91	0,73	0,83
2	75,70	0,76	0,92	0,74	0,84
3	77,41	0,77	0,92	0,75	0,84
4	73,79	0,74	0,91	0,71	0,82
5	77,27	0,77	0,92	0,75	0,84
Promedio	**75,96**	0,76	0,92	0,74	0,83
STD	1,48	0,01	0,01	0,02	0,01

MODELO	SVM				
Repetición	Accuracy	Sensibilidad	Especificidad	F1-Score	AUC-ROC
1	63,07	0,63	0,83	0,51	0,84
2	67,78	0,68	0,83	0,57	0,86
3	63,21	0,63	0,82	0,51	0,84
4	64,03	0,64	0,83	0,52	0,85
5	64,57	0,65	0,83	0,53	0,85
Promedio	**64,53**	0,65	0,83	0,53	0,85
STD	1,92	0,02	0,00	0,02	0,01

MODELO	Random Forest				
Repetición	Accuracy	Sensibilidad	Especificidad	F1-Score	AUC-ROC
1	75,70	0,76	0,90	0,70	0,87
2	72,42	0,72	0,89	0,66	0,87
3	74,06	0,74	0,90	0,68	0,87
4	74,47	0,74	0,90	0,69	0,88
5	75,22	0,75	0,90	0,70	0,87
Promedio	**74,37**	0,74	0,90	0,69	0,87
STD	1,26	0,01	0,00	0,02	0,00

MODELO	XG-Boost				
Repetición	Accuracy	Sensibilidad	Especificidad	F1-Score	AUC-ROC
1	77,34	0,77	0,92	0,75	0,90
2	76,93	0,77	0,92	0,75	0,90
3	74,88	0,75	0,92	0,72	0,90
4	77,00	0,77	0,92	0,75	0,90
5	76,25	0,76	0,92	0,74	0,89
Promedio	**76,48**	0,76	0,92	0,74	0,90
STD	0,98	0,01	0,00	0,01	0,00

MODELO	GBM				
Repetición	Accuracy	Sensibilidad	Especificidad	F1-Score	AUC-ROC
1	74,20	0,74	0,91	0,71	0,89
2	74,27	0,74	0,91	0,71	0,89
3	72,76	0,73	0,91	0,69	0,88
4	72,08	0,72	0,91	0,69	0,89
5	74,74	0,75	0,91	0,72	0,89
Promedio	**73,61**	0,74	0,91	0,70	0,89
STD	1,13	0,01	0,00	0,01	0,00

MODELO	Feed Forward				
Repetición	Accuracy	Sensibilidad	Especificidad	F1-Score	AUC-ROC
1	74,20	0,74	0,93	0,74	0,89
2	73,17	0,73	0,92	0,70	0,90
3	75,29	0,75	0,93	0,73	0,90
4	75,43	0,75	0,92	0,73	0,90
5	74,95	0,75	0,93	0,75	0,90
Promedio	**74,61**	0,74	0,93	0,73	0,90
STD	0,93	0,01	0,01	0,02	0,00

MODELO	RNN				
Repetición	Accuracy	Sensibilidad	Especificidad	F1-Score	AUC-ROC
1	74,33	0,74	0,92	0,72	0,89
2	71,74	0,72	0,92	0,72	0,90
3	73,92	0,74	0,92	0,70	0,89
4	74,88	0,75	0,92	0,73	0,90
5	72,01	0,72	0,93	0,70	0,89
Promedio	**73,38**	0,73	0,92	0,71	0,89
STD	1,42	0,01	0,00	0,01	0,01

MODELO	EnsembleVoting (SVM, FeedForward, RNN)				
Repetición	Accuracy	Sensibilidad	Especificidad	F1-Score	AUC-ROC
1	76,25	0,76	0,92	0,74	0,90
2	74,27	0,74	0,92	0,71	0,90
3	74,13	0,74	0,92	0,71	0,89
4	73,17	0,73	0,92	0,70	0,88
5	76,93	0,77	0,92	0,74	0,90
Promedio	**74,95**	0,75	0,92	0,72	0,89
STD	1,57	0,02	0,00	0,02	0,01

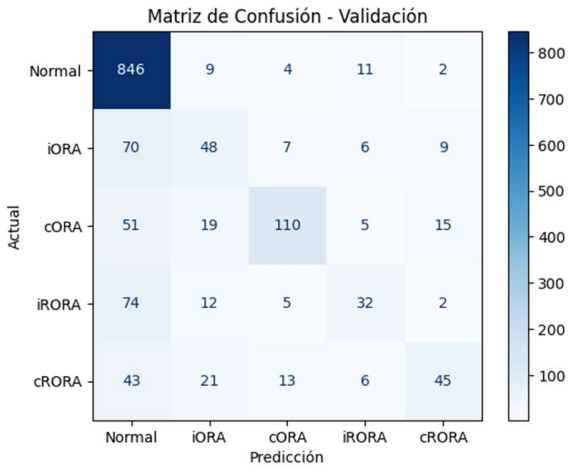
Tabla No. 7 *– Pruebas con VGG16 como extractor de características*

5.3. EXTRACCIÓN DE CARACTERÍTICAS CON ARQUITECTURAS PREENTRENADAS: RESNET50 Y EFFICIENTNET

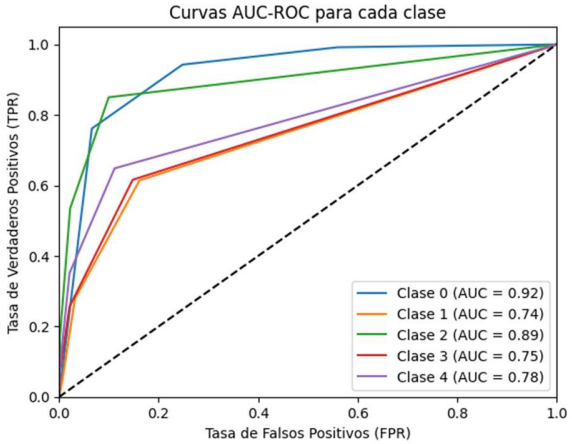
5.3.1 RESNET50 PRE-ENTRENADA + CLASIFICACIÓN CON KNN-3

El modelo K-NN logró una exactitud de 73.60% en promedio (Tabla No. 8) y un AUC-ROC general de 0.82. La clase "*Normal*" mostró un rendimiento sobresaliente con un F1-score de 0.87, debido a su alta precisión (0.78) y recall (0.97). Sin embargo, las clases patológicas como "iORA", "iRORA" y "cRORA" obtuvieron puntajes más bajos, con F1-scores en el rango de 0.35 a 0.45. Esto sugiere que el modelo tiene dificultades para diferenciar estas categorías, posiblemente debido a la superposición visual entre las distintas manifestaciones de la degeneración macular. La matriz de confusión muestra que las clases patológicas se confunden con frecuencia entre sí, mientras que la clase "Normal" se clasifica de manera precisa. La sensibilidad general fue de 0.74. La especificidad general fue alta (0.90), destacando que el modelo tiene una buena habilidad para evitar falsos positivos, especialmente en las clases patológicas.

Figura No. 33 *– Resultados de ResNet50 + Clasificación con KNN-3*



Reporte de Clasificación:	precision	recall	f1-score	support
Normal	0.78	0.97	0.87	872
iORA	0.44	0.34	0.39	140
cORA	0.79	0.55	0.65	200
iRORA	0.53	0.26	0.35	125
cRORA	0.62	0.35	0.45	128
accuracy			0.74	1465
macro avg	0.63	0.49	0.54	1465
weighted avg	0.71	0.74	0.71	1465



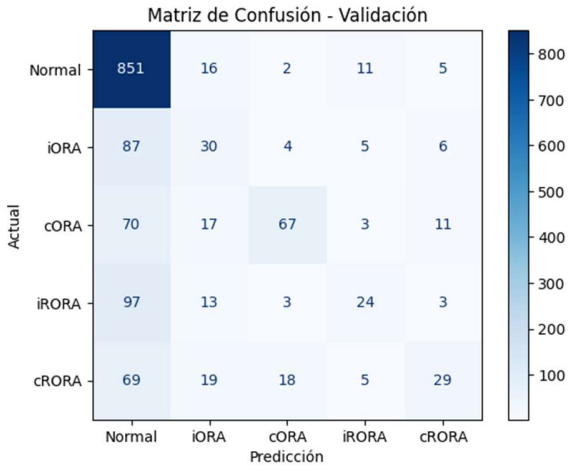
<u>Métricas modelo KNN-3</u>:

Exactitud del modelo K-NN: 73.79%
AUC-ROC General: 0.82
F1-Score Ponderado: 0.71
Sensibilidad General: 0.74
Especificidad Clase 1: 0.60
Especificidad Clase 2: 0.95
Especificidad Clase 3: 0.98
Especificidad Clase 4: 0.98
Especificidad Clase 5: 0.98
Especificidad General: 0.90

5.3.2. EFFICIENTNET PRE-ENTRENADA Y CLASIFICACIÓN CON KNN-3

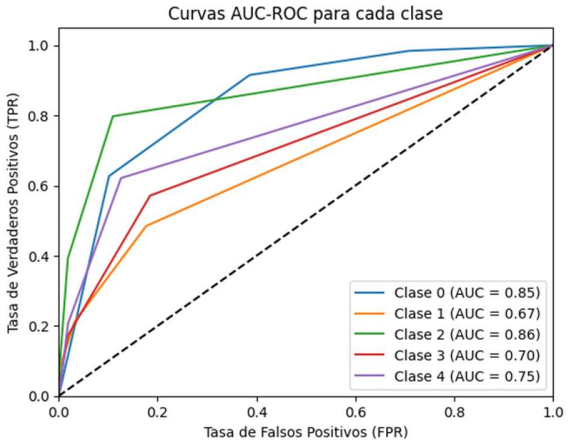
El rendimiento del modelo muestra una exactitud promedio del 68.59% (Tabla No. 8), con un AUC-ROC general de 0.77 y un F1-Score ponderado de 0.63. La clase 'Normal' obtuvo los mejores resultados, con un F1-score de 0.83, debido a una alta precisión (0.72) y recall (0.96. Sin embargo, las clases patológicas como 'iORA', 'iRORA' y 'cRORA' presentan F1-scores más bajos, que varían entre 0.26 y 0.51, lo que refleja una menor capacidad del modelo para distinguir entre estas categorías. La matriz de confusión muestra una alta proporción de predicciones correctas para la clase 'Normal', pero indica confusiones significativas entre las clases patológicas. Por ejemplo, 'iORA' y 'cORA' se clasifican incorrectamente con frecuencia. La sensibilidad general fue de 0.68, lo que indica que el modelo tiene una capacidad moderada para identificar verdaderos positivos, mientras que la especificidad general fue de 0.87, destacando una mejor habilidad para evitar falsos positivos.

Figura No. 34 *– Resultados de EfficientNet + Clasificación con KNN-3*



Reporte de Clasificación:

	precision	recall	f1-score	support
Normal	0.72	0.96	0.83	885
iORA	0.32	0.23	0.26	132
cORA	0.71	0.40	0.51	168
iRORA	0.50	0.17	0.26	140
cRORA	0.54	0.21	0.30	140
accuracy			0.68	1465
macro avg	0.56	0.39	0.43	1465
weighted avg	0.65	0.68	0.63	1465



<u>Métricas modelo KNN-3</u>:

Exactitud del modelo K-NN (k=3): 68.33%
AUC-ROC General: 0.77
F1-Score Ponderado: 0.63
Sensibilidad General: 0.68
Especificidad Clase 1: 0.44
Especificidad Clase 2: 0.95
Especificidad Clase 3: 0.98
Especificidad Clase 4: 0.98
Especificidad Clase 5: 0.98
Especificidad General: 0.87

Tabla No. 8 – *Comparativo entre las repeticiones de los modelos realizados de las arquitecturas ResNet50 y EfficientNet con el clasificador KNN-3.*

ResNet: Extractor de características					
MODELO	**K-NN (3)**				
Repetición	**Accuracy**	**Sensibilidad**	**Especificidad**	**F1-Score**	**AUC-ROC**
1	75,22	0,75	0,90	0,73	0,82
2	71,67	0,72	0,89	0,68	0,80
3	73,52	0,74	0,90	0,71	0,82
4	73,86	0,74	0,90	0,71	0,81
5	73,72	0,74	0,90	0,71	0,81
Promedio	**73,60**	0,74	0,90	0,71	0,81
STD	1,27	0,01	0,00	0,02	0,01

EfficientNet: Extractor de características					
MODELO	**K-NN (3)**				
Repetición	**Accuracy**	**Sensibilidad**	**Especificidad**	**F1-Score**	**AUC-ROC**
1	68,60	0,69	0,87	0,64	0,76
2	69,49	0,69	0,87	0,65	0,76
3	67,99	0,68	0,86	0,63	0,76
4	68,53	0,69	0,87	0,64	0,76
5	68,33	0,68	0,87	0,63	0,77
Promedio	**68,59**	0,69	0,87	0,64	0,76
STD	0,56	0,01	0,00	0,01	0,00

5.4. DISCUSIÓN

Se validaron las diferentes arquitecturas como extractoras de características, con sus respectivos clasificadores en 5 repeticiones cada una, para evidenciar su reproducibilidad. A continuación, se presenta una discusión sobre los resultados obtenidos, enfocándose especialmente en la exactitud alcanzada por cada arquitectura + clasificador, así como en las métricas relevantes descritas en el texto.

5.4.1. MODELO CON CNN PROPIA PRE-ENTRENADA

El modelo híbrido que combinó una CNN propia pre-entrenada como extractor de características y un clasificador SVM obtuvo una exactitud del 88.60% (87.80% en promedio), siendo el más alto entre todos los modelos probados. Además, mostró un AUC-ROC general de 0.94 y un F1-Score ponderado de 0.88, indicando una excelente capacidad para distinguir entre las diferentes clases. La alta especificidad general del 97% sugiere que el modelo es efectivo para evitar falsos positivos, mientras que una sensibilidad general del 89% refleja una sólida detección de verdaderos positivos.

Utilizando KNN con k=3, el modelo alcanzó una exactitud del 83.62% y un AUC-ROC de 0.89. Aunque ligeramente inferior al modelo SVM, mantiene un buen desempeño general. El F1-Score ponderado fue de 0.83, y la especificidad general fue del 95%, indicando una buena precisión en las predicciones. Sin embargo, la sensibilidad general fue del 84%, sugiriendo que el modelo podría mejorar en la detección de algunas clases minoritarias.

Con Random Forest, el modelo logró una exactitud del 84.37% y un AUC-ROC de 0.95. El F1-Score ponderado fue de 0.83, similar al obtenido con KNN-3. La especificidad general se mantuvo alta en 95%, y la sensibilidad general fue del 84%. Esto indica que el modelo es consistente en la clasificación correcta, aunque con margen para mejorar en la sensibilidad.

El uso de XGBoost resultó en una exactitud del 86.28% y un AUC-ROC de 0.96, mostrando una mejora respecto a Random Forest y KNN-3. El F1-Score ponderado aumentó a 0.86, y tanto la especificidad como la sensibilidad, generales, fueron del 96% y 86% respectivamente, evidenciando un equilibrio entre la detección de verdaderos positivos y la reducción de falsos positivos.

El modelo con Gradient Boosting Machines alcanzó una exactitud del 85.60% y un AUC-ROC de 0.95. El F1-Score ponderado fue de 0.85, con una especificidad general del 96% y una sensibilidad general del 86%. Aunque similar a XGBoost, presenta una ligera disminución en la exactitud y el F1-Score.

Este modelo obtuvo una exactitud del 87.24% y un AUC-ROC de 0.96, posicionándose como uno de los modelos con mejor rendimiento. El F1-Score ponderado fue de 0.87, con una especificidad

general del 96% y una sensibilidad general del 87%. Estos resultados indican una sólida capacidad para clasificar correctamente las imágenes en las diferentes clases.

La implementación de una Red Neuronal Recurrente resultó en una exactitud del 87.44% y un AUC-ROC de 0.97, ligeramente superior al modelo Feedforward. El F1-Score ponderado también fue de 0.87, con especificidad y sensibilidad generales de 96% y 87% respectivamente. Este modelo demuestra una alta capacidad discriminativa, especialmente reflejada en el AUC-ROC.

Al combinar los modelos SVM, Feedforward y RNN mediante Ensemble Voting, se obtuvo una exactitud del 85.32% y un AUC-ROC de 0.96. El F1-Score ponderado fue de 0.84, con una especificidad general del 95% y una sensibilidad general del 85%. Aunque el ensamblaje buscaba mejorar el rendimiento general, los resultados indican que no superó a los modelos individuales de Feedforward y RNN.

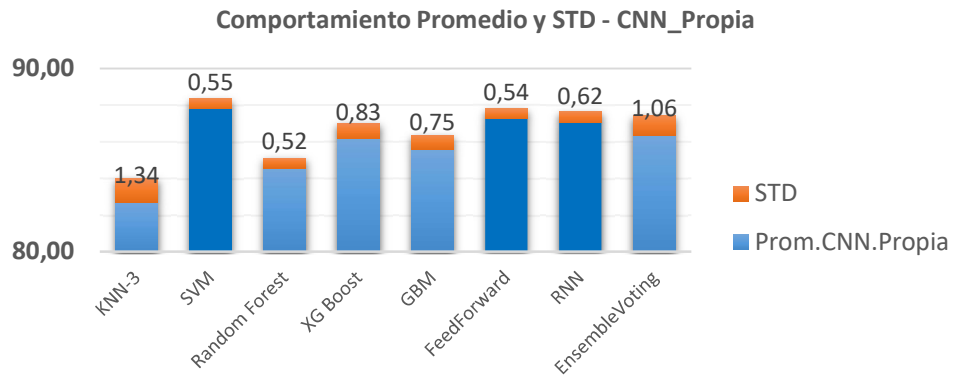


***Gráfico No. 1 – Comportamiento del promedio y la STD de la CNN Propia**. Se puede validar como las pruebas presentan una mayor exactitud para SVM, no obstante, seguidas de FeedFrwd y RNN.*

5.4.2. MODELO CON VGG16 PRE-ENTRENADA

Al utilizar la arquitectura VGG16 como extractor de características, se observó una disminución en la exactitud y el rendimiento general en comparación con la CNN propia.

Clasificación con SVM: la exactitud alcanzada fue del 63.14%, con un AUC-ROC de 0.85 y un F1-Score ponderado de 0.51. Estos resultados indican un desempeño limitado, con dificultades para clasificar correctamente las clases patológicas.

Clasificación con KNN-3: El modelo obtuvo una exactitud del 75.63% y un AUC-ROC de 0.83. El F1-Score ponderado fue de 0.73, mostrando una mejora respecto al SVM, pero aún por debajo de los modelos basados en la CNN propia.

Clasificación con Random Forest y XGBoost: Con Random Forest, la exactitud fue del 75.22%, mientras que con XGBoost se alcanzó el 76.25%. Los AUC-ROC fueron de 0.87 y 0.89 respectivamente. Aunque estos modelos mejoraron ligeramente el rendimiento, no lograron alcanzar los niveles de exactitud de los modelos con la CNN propia.

Clasificación con GBM, Feedforward y RNN: Los modelos GBM y Feedforward presentaron una exactitud cercana al 75%, mientras que el modelo RNN obtuvo un 72.01%. Los AUC-ROC oscilaron entre 0.89 y 0.90. Estos resultados refuerzan la observación de que VGG16 no proporcionó características tan discriminativas como la CNN propia para este conjunto de datos.

Clasificación con Ensemble Voting: El ensamblaje de modelos con VGG16 alcanzó una exactitud del 76.93% y un AUC-ROC de 0.90. Aunque hubo una ligera mejora, el rendimiento sigue siendo inferior al obtenido con la CNN propia.

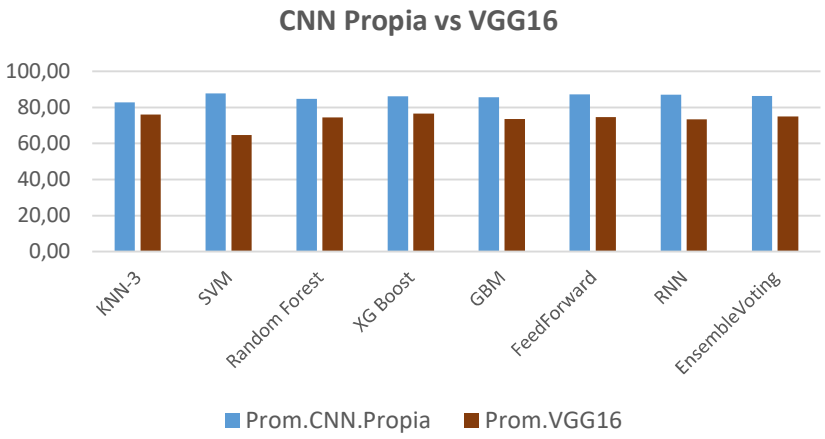


Gráfico No. 2 – Comparativo entre la CNN Propia y VGG16. *Se observa como las múltiples repeticiones realizadas para cada uno de los clasificadores, presentan mejores resultados en la CNN Propia que en la VGG16.*

5.4.3. MODELO CON RESNET50 Y EFFICIENTNET PRE-ENTRENADAS

Debido al alto costo computacional y al tamaño de los archivos de características (4.2 GB para ResNet50 y 2.08 GB para EfficientNet), se realizaron pruebas limitadas utilizando KNN-3.

Clasificación con KNN-3 y ResNet50: El modelo logró una exactitud del 73.79% y un AUC-ROC de 0.82. Estos resultados son similares a los obtenidos con VGG16, indicando un rendimiento moderado.

Clasificación con KNN-3 y EfficientNet: Se obtuvo una exactitud del 68.33% y un AUC-ROC de 0.77, siendo este el rendimiento más bajo entre los modelos probados. Esto sugiere que, para este conjunto de datos y con KNN-3, EfficientNet no proporcionó mejoras significativas.

5.5 ANÁLISIS GENERAL

Los resultados indican que la CNN propia pre-entrenada, combinada con clasificadores como SVM, Feedforward Neural Network y RNN, ofreció los mejores rendimientos en términos de exactitud y métricas asociadas. La exactitud más alta se obtuvo con el SVM (88.60%), seguido de cerca por los modelos Feedforward y RNN con 87.24% y 87.44% respectivamente (Gráfico No. 1). Estos modelos también mostraron altos valores de AUC-ROC y F1-Score ponderado, indicando una excelente capacidad para distinguir entre las diferentes clases y un buen equilibrio entre precisión y sensibilidad.

Por otro lado, los modelos que utilizaron VGG16, ResNet50 y EfficientNet como extractores de características no alcanzaron el mismo nivel de rendimiento. Esto podría deberse a que la CNN propia fue específicamente diseñada y ajustada para el conjunto de datos y las características particulares de las imágenes OCT relacionadas con la DMAE, capturando patrones más relevantes para la clasificación.

Es importante destacar que, aunque los modelos de ensamblaje (Ensemble Voting) buscaban combinar las fortalezas de diferentes clasificadores, no lograron superar el rendimiento de los mejores modelos individuales. Esto sugiere que, en este caso, un modelo bien ajustado y especializado es más efectivo que la combinación de varios modelos con rendimientos dispares.

En conclusión, la elección de la arquitectura de la CNN y del clasificador tuvo un impacto significativo en el rendimiento del modelo. Los resultados evidenciaron que una CNN propia, diseñada y entrenada específicamente para el problema en cuestión, combinada con clasificadores robustos como SVM, Feedforward Neural Networks o RNN, ofrece la mejor exactitud y capacidad de generalización para la estadificación imagenológica de la DMAE a partir de imágenes OCT.

6. HALLAZGOS

6.1 IMPLICACIONES CLÍNICAS Y TÉCNICAS

Los hallazgos demuestran que la elección de la arquitectura de la CNN y del clasificador tiene un impacto significativo en el rendimiento del modelo. La capacidad de la CNN propia para extraer características más discriminativas se traduce en una mejor clasificación de las etapas de la DMAE. Esto es crucial en el contexto clínico, ya que una clasificación precisa puede ayudar a los oftalmólogos en el diagnóstico temprano y el seguimiento de la enfermedad.

Además, la alta especificidad y sensibilidad alcanzadas por los mejores modelos indican una baja tasa de falsos positivos y falsos negativos, lo cual es esencial para minimizar errores en diagnósticos médicos [31,41].

6.2 LIMITACIONES

Una de las limitaciones encontradas fue el alto costo computacional asociado con modelos más complejos y el procesamiento de grandes volúmenes de datos [42]. Aunque las arquitecturas preentrenadas como ResNet50 y EfficientNetB0 son poderosas, su implementación requiere recursos significativos, lo que puede no ser práctico en entornos con limitaciones de hardware y banda de internet.

7. RECOMENDACIONES

7.1 INTEGRACIÓN DEL MODELO EN LA PRÁCTICA CLÍNICA DE LA CLÍNICA VISUAL Y AUDITIVA

Se recomienda implementar el modelo de clasificación imagenológica desarrollado en este proyecto directamente en el flujo de trabajo de la Clínica Visual y Auditiva del Instituto para Niños Ciegos y Sordos. Esto permitiría a los oftalmólogos utilizar la herramienta como apoyo en el diagnóstico y estadificación de la Degeneración Macular Atrófica Relacionada con la Edad a partir de imágenes de Tomografía de Coherencia Óptica. Esta integración del modelo puede mejorar la precisión diagnóstica, facilitar la detección temprana de la enfermedad y optimizar los planes de tratamiento, beneficiando directamente a los pacientes atendidos en la clínica.

7.2 ACTUALIZACIÓN Y MANTENIMIENTO CONTINUO DEL MODELO CON NUEVOS DATOS

Se recomienda establecer un proceso continuo de actualización utilizando nuevas imágenes de OCT recolectadas en la clínica. Esto implica desarrollar un sistema para incorporar datos nuevos y variados, permitiendo que el modelo aprenda de casos recientes y se adapte a posibles cambios en las características de la población atendida.

7.3 AMPLIACIÓN DEL CONJUNTO DE DATOS CON PACIENTES DE DIVERSAS CARACTERÍSTICAS

Se sugiere ampliar el conjunto de datos incluyendo imágenes de pacientes con diferentes características demográficas y clínicas, como diversas edades, etnias y grados de progresión de la DMAE. Esto mejorará la capacidad de generalización del modelo y su aplicabilidad a una población más amplia.

7.4 COLABORACIÓN INTERINSTITUCIONAL PARA EL DESARROLLO TECNOLÓGICO Y COMPARTICIÓN DE CONOCIMIENTOS

Fomentar colaboraciones con otras instituciones a través de alianzas estratégicas, con las que el Instituto pueda compartir experiencias, datos anonimizados y mejores prácticas, lo que contribuirá al mejoramiento del modelo y al desarrollo de nuevas soluciones innovadoras en esta materia.

8. CONCLUSIONES Y TRABAJOS FUTUROS

8.1 CONCLUSIONES

Los resultados obtenidos evidencian que una CNN propia, diseñada y entrenada específicamente para el conjunto de datos y el problema de clasificación de la degeneración macular, combinada con clasificadores robustos como SVM, Redes Neuronales Feedforward o RNN, ofrece la mejor exactitud y capacidad de generalización. Estos modelos superan a las arquitecturas preentrenadas estándar, destacando la importancia de adaptar y personalizar los modelos a las características específicas de los datos médicos utilizados.

Este estudio contribuye al avance en el diagnóstico y seguimiento de la DMAE mediante técnicas de aprendizaje automático realizadas en población colombiana, proporcionando una herramienta potencialmente valiosa para apoyar la toma de decisiones clínicas y mejorar los resultados para nuestros pacientes.

8.2 TRABAJOS FUTUROS

Para trabajos futuros, se propone profundizar en varias áreas para mejorar y ampliar los hallazgos de este estudio. En primer lugar, la recolección de un conjunto de datos más amplio y diverso podría ayudar a mejorar la capacidad de generalización del modelo, especialmente en lo que respecta a las clases menos representadas. La inclusión de imágenes de OCT provenientes de diferentes dispositivos y de pacientes con diversas características demográficas aumentaría la robustez del modelo y su aplicabilidad en distintos contextos clínicos. Además, la implementación de técnicas de aprendizaje transferencial más avanzadas y la exploración de arquitecturas de redes neuronales más eficientes, como MobileNet o NASNet, podrían reducir el costo computacional y facilitar el despliegue del modelo en entornos con recursos limitados [50,60].

También sería valioso investigar la integración de técnicas de interpretación de modelos para mejorar la interpretabilidad y transparencia del sistema, lo cual es crucial en aplicaciones médicas [62].

Finalmente, se recomienda llevar a cabo una validación clínica del modelo en colaboración con profesionales de la salud, evaluando su desempeño en entornos reales y su impacto en la toma de decisiones médicas, lo que sería un paso fundamental para la adopción de esta tecnología en la práctica clínica.

9. GLOSARIO

Accuracy (Precisión/Exactitud): Medida de rendimiento de un modelo que indica el porcentaje de predicciones correctas sobre el total de casos evaluados.

Adam (Adaptive Moment Estimation): Optimizador utilizado en el entrenamiento de redes neuronales que combina las ventajas de AdaGrad y RMSProp para adaptar la tasa de aprendizaje de cada parámetro.

Ajuste de Hiperparámetros: Proceso de selección y ajuste de los parámetros configurables de un modelo de aprendizaje automático para optimizar su rendimiento.

Área Bajo la Curva ROC (AUC-ROC): Métrica que resume la capacidad de un modelo para distinguir entre clases, calculando el área bajo la curva ROC (Receiver Operating Characteristic).

Arquitectura de Red: Diseño estructural de una red neuronal, incluyendo el número y tipo de capas, neuronas y conexiones entre ellas.

Atrofia Geográfica (AG): Etapa avanzada de la Degeneración Macular Atrófica Relacionada con la Edad caracterizada por la pérdida de células del epitelio pigmentario de la retina.

Capa Convolucional: Componente básico de las redes neuronales convolucionales que aplica filtros para extraer características locales de las imágenes.

Capa Densa (Fully Connected Layer): Capa en una red neuronal donde cada neurona está conectada a todas las neuronas de la capa anterior, utilizada para combinar características y realizar la clasificación final.

CNN (Red Neuronal Convolucional): Tipo de red neuronal diseñada para procesar datos con estructura de cuadrícula, como imágenes, y extraer características jerárquicas.

Conjunto de Datos de Entrenamiento: Subconjunto de datos utilizado para ajustar los parámetros del modelo durante el entrenamiento.

Conjunto de Datos de Validación: Subconjunto de datos utilizado para evaluar el rendimiento del modelo durante el entrenamiento y ajustar hiperparámetros.

Curva ROC (Receiver Operating Characteristic): Gráfico que representa la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos a diferentes umbrales de clasificación.

Data Augmentation (Aumento de Datos): Técnicas que generan nuevas muestras de datos a partir de las existentes mediante transformaciones como rotaciones, desplazamientos o cambios de

escala para aumentar la diversidad del conjunto de datos.

Degeneración Macular Atrófica Relacionada con la Edad (DMAE): Enfermedad ocular que afecta la mácula y conduce a la pérdida progresiva de la visión central en personas mayores.

Desbalance de Clases: Situación en la que las clases en un conjunto de datos no están representadas de manera equitativa, lo que puede afectar el rendimiento del modelo.

Dropout: Técnica de regularización que desactiva aleatoriamente neuronas durante el entrenamiento para prevenir el sobreajuste.

Early Stopping (Detención Temprana): Método que detiene el entrenamiento de un modelo cuando el rendimiento en el conjunto de validación deja de mejorar, evitando el sobreajuste.

Ecualización del Histograma: Técnica de procesamiento de imágenes que mejora el contraste ajustando la distribución de los niveles de intensidad.

EfficientNet: Familia de arquitecturas de redes neuronales convolucionales optimizadas para lograr un equilibrio entre rendimiento y eficiencia computacional.

Hiperparámetros: Parámetros configurables de un modelo de aprendizaje automático que deben ser establecidos antes del entrenamiento y que afectan su comportamiento y rendimiento.

Keras: Biblioteca de código abierto para el desarrollo de redes neuronales en Python, que funciona como una interfaz de alto nivel para TensorFlow.

Normalización: Proceso de ajuste de los valores de datos a una escala común para mejorar la eficiencia del entrenamiento y la estabilidad del modelo.

OCT (Tomografía de Coherencia Óptica): Técnica de imagen médica que utiliza luz de baja coherencia para capturar imágenes de alta resolución de las estructuras internas del ojo.

Optimizador: Algoritmo que ajusta los pesos y sesgos del modelo durante el entrenamiento para minimizar la función de pérdida.

Overfitting (Sobreajuste): Situación en la que un modelo se ajusta demasiado a los datos de entrenamiento y falla al generalizar a nuevos datos.

Pesos por Clase: Valores asignados a cada clase durante el entrenamiento para manejar el desbalance de clases, dando más importancia a las clases minoritarias.

Regularización: Técnicas utilizadas para prevenir el sobreajuste añadiendo información adicional o restricciones al modelo.

ResNet50: Arquitectura de red neuronal profunda que introduce conexiones residuales para permitir el entrenamiento de redes más profundas sin problemas de degradación.

TensorFlow: Biblioteca de código abierto para el cálculo numérico y aprendizaje automático desarrollada por Google.

Validación Cruzada (CrossValidation): Técnica para evaluar el rendimiento de un modelo dividiendo el conjunto de datos en múltiples subconjuntos de entrenamiento y validación.

Validación Cruzada kFold: Variante de la validación cruzada donde el conjunto de datos se divide en k grupos, y el modelo se entrena y valida k veces, cada vez con un grupo diferente como conjunto de validación.

VGG16: Arquitectura de red neuronal convolucional profunda con 16 capas, conocida por su simplicidad y eficacia en tareas de clasificación de imágenes.

10. REFERENCIAS BIBLIOGRÁFICAS

1. Thomas CJ, Mirza RG, Gill MK. Age-Related Macular Degeneration. Med Clin North Am. 2021 May;105(3):473-491. doi: 10.1016/j.mcna.2021.01.003. Epub 2021 Apr 2. PMID: 33926642.
2. Gheorghe A, Mahdi L, Musat O. AGE-RELATED MACULAR DEGENERATION. Rom J Ophthalmol. 2015 Apr-Jun;59(2):74-7. PMID: 26978865; PMCID: PMC5712933.
3. Mitchell P, Liew G, Gopinath B, Wong TY. Age-related macular degeneration. Lancet. 2018 Sep 29;392(10153):1147-1159. doi: 10.1016/S0140-6736(18)31550-2. PMID: 30303083.
4. Chee RI, Mahrous A, Koenig L, Mandel LS, Yazdanie F, Chan CC, Gupta MP. Histopathology of Age-Related Macular Degeneration and Implications for Pathogenesis and Therapy. Adv Exp Med Biol. 2021;1256:67-88. doi: 10.1007/978-3-030-66014-7_3. PMID: 33847998.
5. Zhang G, Fu DJ, Liefers B, Faes L, Glinton S, Wagner S, Struyven R, Pontikos N, Keane PA, Balaskas K. Clinically relevant deep learning for detection and quantification of geographic atrophy from optical coherence tomography: a model development and external validation study. Lancet Digit Health. 2021 Oct;3(10):e665-e675. doi: 10.1016/S2589-7500(21)00134-5. Epub 2021 Sep 8. PMID: 34509423.
6. Balaskas K, Glinton S, Keenan TDL, Faes L, Liefers B, Zhang G, Pontikos N, Struyven R, Wagner SK, McKeown A, Patel PJ, Keane PA, Fu DJ. Prediction of visual function from automatically quantified optical coherence tomography biomarkers in patients with geographic atrophy using machine learning. Sci Rep. 2022 Sep 16;12(1):15565. doi: 10.1038/s41598-022-19413-z. PMID: 36114218; PMCID: PMC9481631.
7. Keenan TDL, Cukras CA, Chew EY. Age-Related Macular Degeneration: Epidemiology and Clinical Aspects. Adv Exp Med Biol. 2021;1256:1-31. doi: 10.1007/978-3-030-66014-7_1. PMID: 33847996.
8. Waldstein SM, Vogl WD, Bogunovic H, Sadeghipour A, Riedl S, Schmidt-Erfurth U. Characterization of Drusen and Hyperreflective Foci as Biomarkers for Disease Progression in Age-Related Macular Degeneration Using Artificial Intelligence in Optical Coherence Tomography. JAMA Ophthalmol. 2020 Jul 1;138(7):740-747. doi: 10.1001/jamaophthalmol.2020.1376. PMID: 32379287; PMCID: PMC7206537.
9. Romond K, Alam M, Kravets S, Sisternes L, Leng T, Lim JI, Rubin D, Hallak JA. Imaging and artificial intelligence for progression of age-related macular degeneration. Exp Biol Med (Maywood). 2021 Oct;246(20):2159-2169. doi: 10.1177/15353702211031547. Epub 2021 Aug 18. PMID: 34404252; PMCID: PMC8718252.
10. Bogunovic H, Montuoro A, Baratsits M, Karantonis MG, Waldstein SM, Schlanitz F, Schmidt-Erfurth U. Machine Learning of the Progression of Intermediate Age-Related Macular

Degeneration Based on OCT Imaging. Invest Ophthalmol Vis Sci. 2017 May 1;58(6):BIO141-BIO150. doi: 10.1167/iovs.17-21789. PMID: 28658477.

11. Kalra G, Cetin H, Whitney J, Yordi S, Cakir Y, McConville C, Whitmore V, Bonnay M, Lunasco L, Sassine A, Borisiak K, Cohen D, Reese J, Srivastava SK, Ehlers JP. Machine Learning-Based Automated Detection and Quantification of Geographic Atrophy and Hypertransmission Defects Using Spectral Domain Optical Coherence Tomography. J Pers Med. 2022 Dec 24;13(1):37. doi: 10.3390/jpm13010037. PMID: 36675697; PMCID: PMC9861976.
12. Lin AC, Lee CS, Blazes M, Lee AY, Gorin MB. Assessing the Clinical Utility of Expanded Macular OCTs Using Machine Learning. Transl Vis Sci Technol. 2021 May 3;10(6):32. doi: 10.1167/tvst.10.6.32. PMID: 34038502; PMCID: PMC8161701.
13. Boyer DS, Schmidt-Erfurth U, van Lookeren Campagne M, Henry EC, Brittain C. THE PATHOPHYSIOLOGY OF GEOGRAPHIC ATROPHY SECONDARY TO AGE-RELATED MACULAR DEGENERATION AND THE COMPLEMENT PATHWAY AS A THERAPEUTIC TARGET. Retina. 2017 May;37(5):819-835. doi: 10.1097/IAE.0000000000001392. PMID: 27902638; PMCID: PMC5424580.
14. Ung C, Lains I, Miller JW, Kim IK. Current Management of Age-Related Macular Degeneration. Adv Exp Med Biol. 2021;1256:295-314. doi: 10.1007/978-3-030-66014-7_12. PMID: 33848007.
15. Wright CB, Ambati J. Dry Age-Related Macular Degeneration Pharmacology. Handb Exp Pharmacol. 2017;242:321-336. doi: 10.1007/164_2016_36. PMID: 27900609; PMCID: PMC5472449.
16. Treder M, Lauermann JL, Eter N. Deep learning-based detection and classification of geographic atrophy using a deep convolutional neural network classifier. Graefes Arch Clin Exp Ophthalmol. 2018 Nov;256(11):2053-2060. doi: 10.1007/s00417-018-4098-2. Epub 2018 Aug 8. PMID: 30091055.
17. Shi X, Keenan TDL, Chen Q, De Silva T, Thavikulwat AT, Broadhead G, Bhandari S, Cukras C, Chew EY, Lu Z. Improving Interpretability in Machine Diagnosis: Detection of Geographic Atrophy in OCT Scans. Ophthalmol Sci. 2021 Jul 13;1(3):100038. doi: 10.1016/j.xops.2021.100038. PMID: 36247813; PMCID: PMC9559084.
18. Gigon A, Mosinska A, Montesel A, Derradji Y, Apostolopoulos S, Ciller C, De Zanet S, Mantel I. Personalized Atrophy Risk Mapping in Age-Related Macular Degeneration. Transl Vis Sci Technol. 2021 Nov 1;10(13):18. doi: 10.1167/tvst.10.13.18. PMID: 34767623; PMCID: PMC8590159.
19. Schmidt-Erfurth U, Waldstein SM, Klimscha S, Sadeghipour A, Hu X, Gerendas BS, Osborne A, Bogunovic H. Prediction of Individual Disease Conversion in Early AMD Using Artificial Intelligence. Invest Ophthalmol Vis Sci. 2018 Jul 2;59(8):3199-3208. doi: 10.1167/iovs.18-24106. PMID: 29971444.

20. Derradji Y, Mosinska A, Apostolopoulos S, Ciller C, De Zanet S, Mantel I. Fully-automated atrophy segmentation in dry age-related macular degeneration in optical coherence tomography. Sci Rep. 2021 Nov 8;11(1):21893. doi: 10.1038/s41598-021-01227-0. PMID: 34751189; PMCID: PMC8575929.
21. Feeny AK, Tadarati M, Freund DE, Bressler NM, Burlina P. Automated segmentation of geographic atrophy of the retinal epithelium via random forests in AREDS color fundus images. Comput Biol Med. 2015 Oct 1;65:124-36. doi: 10.1016/j.compbiomed.2015.06.018. Epub 2015 Jul 9. PMID: 26318113; PMCID: PMC4670087.
22. Sadda, S. R., Guymer, R., Holz, F. G., Schmitz-Valckenberg, S., Curcio, C. A., et al. (2018). Consensus Definition for Atrophy Associated with Age-Related Macular Degeneration on OCT: Classification of Atrophy Report 3. Ophthalmology, 125(4), 537-548. doi:10.1016/j.ophtha.2017.09.028.
23. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436-444, 2015.
24. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.
25. J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Networks, vol. 61, pp. 85-117, 2015.
26. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2012, vol 25.
27. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proceedings of the International Conference on Learning Representations (ICLR), 2015.
28. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, pp. 234-241.
29. Fujimoto JG, Swanson EA, Huang D. Optical Coherence Tomography—History, Evolution, and Future Prospects: 2023 Lasker-DeBakey Clinical Medical Research Award. JAMA. 2023;330(15):1427–1428. doi:10.1001/jama.2023.16942
30. S. Tzaridis, et al. Optical coherence tomography: when a picture is worth a million words. 2023;133(19): e174951. https://doi.org/10.1172/JCI174951.
31. D. Shen, G. Wu, and H. Suk, "Deep learning in medical image analysis," Annual Review of Biomedical Engineering, vol. 19, pp. 221-248, 2017.
32. R. A. Leitgeb, "Optical coherence tomography," in Handbook of Coherent-Domain Optical Methods, Springer, 2013, pp. 101-124.

33. J. G. Fujimoto and E. A. Swanson, "The development, commercialization, and impact of optical coherence tomography," Investigative Ophthalmology & Visual Science, vol. 57, no. 9, pp. OCT1-OCT13, 2016.
34. J. M. Schmitt, "Optical coherence tomography (OCT): A review," IEEE Journal of Selected Topics in Quantum Electronics, vol. 5, no. 4, pp. 1205-1215, 1999.
35. T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861-874, 2006.
36. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.
37. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.
38. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21–27.
39. Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874.
40. Gonzalez, R. C., & Woods, R. E. (2002). Digital Image Processing. Prentice Hall.
41. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
42. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284.
43. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. En Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770–778).
44. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. En Proceedings of the 14th International Joint Conference on Artificial Intelligence (Vol. 2, pp. 1137–1143).
45. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. En Advances in Neural Information Processing Systems (pp. 1097–1105).
46. Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., García, D., ... & Korthikanti, V. (2018). Mixed precision training. En International Conference on Learning Representations.
47. Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345–1359.
48. Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. En International Conference on Learning Representations.
49. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437.

50. Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. En Proceedings of the 36th International Conference on Machine Learning (pp. 6105–6114).
51. Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? En Advances in Neural Information Processing Systems (pp. 3320–3328).
52. N. A. Adams. (3013). Atlas of OCT, Retinal Anatomy in Health & Pathology. Heidelberg, Germany: Heidelberg Engineering, Inc.
53. G. Staurenghi, M.D. et al. (2014). For the International Nomenclature for Optical Coherence Tomography (IN-OCT) Panel, "Proposed lexicon for anatomic landmarks in normal posterior segment spectral-domain optical coherence tomography: The IN-OCT consensus," Ophthalmology, vol. 121, pp. 1572-1578, 2014. American Academy of Ophthalmology.
54. B. Villanueva. (2017). Análisis de imágenes oftalmológicas de Tomografía por Coherencia Óptica OCT. Trabajo Fin de Máster presentado en la Escuela Técnica Superior de Ingenieros de Telecomunicación de la Universitat Politècnica de València, para la obtención del Título de Máster en Ingeniería de Telecomunicación. Valencia, España: Universidad Politécnica de Valencia.
55. L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
56. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785–794.
57. J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
58. Z. C. Lipton, J. Berkowitz, and C. Elkan, "A Critical Review of Recurrent Neural Networks for Sequence Learning," arXiv preprint arXiv:1506.00019, 2015.
59. D. Opitz and R. Maclin, "Popular Ensemble Methods: An Empirical Study," Journal of Artificial Intelligence Research, vol. 11, pp. 169–198, 1999.