Universidad Politécnica de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Grado en Ciencia de Datos e Inteligencia Artificial

Trabajo Fin de Grado

Detección Automatizada de Enfermedades Visuales mediante Deep Learning

Autor: Mario Ruiz Vaquett
Tutor(a): Ángel Mario García Pedrero

Madrid, Junio 2025

Este Trabajo Fin de Grado se ha depositado en la ETSI Informáticos de la Universidad Politécnica de Madrid para su defensa.

Trabajo Fin de Grado
Grado en Ciencia de Datos e Inteligencia Artificial

Título: Detección Automatizada de Enfermedades Visuales mediante Deep Learning

Junio 2025

Autor: Mario Ruiz Vaquett
Tutor: Ángel Mario García Pedrero
Departamento de Arquitectura y Tecnología de Sistemas Informáticos
Escuela Técnica Superior de Ingenieros Informáticos
Universidad Politécnica de Madrid

Tabla de contenidos

Resumen

El uso de imágenes médicas como las tomografías de coherencia óptica (OCT) y las fotografías del fondo de ojo (fundus) resulta fundamental en el diagnóstico y seguimiento de enfermedades oculares. Estas modalidades ofrecen información complementaria: mientras las imágenes fundus permiten visualizar la superficie de la retina, las OCT proporcionan una vista transversal que revela detalles estructurales en profundidad. Sin embargo, interpretar estas imágenes de forma precisa requiere experiencia clínica, y su análisis de forma manual puede ser propenso a errores.

La aplicación de técnicas de inteligencia artificial puede contribuir significativamente a la automatización del diagnóstico médico, ofreciendo apoyo a los especialistas, mejorando la eficiencia del proceso y facilitando el acceso a un diagnóstico asistido en centros con recursos limitados. En este contexto, la combinación de información multimodal a través de algoritmos de aprendizaje profundo representa una línea de investigación prometedora.

Este Trabajo de Fin de Grado propone como *baseline* que la combinación de imágenes fundus y OCT, mediante estrategias de fusión de características, permite mejorar el rendimiento de los modelos frente al uso individual de cada modalidad. Para ello, se exploran diferentes técnicas de fusión (*early*, *intermediate* y *late fusion*) empleando como extractores de características los modelos DINOv2 y CLIP, entrenados sobre un conjunto de datos de imágenes fundus y OCT.

La metodología desarrollada abarca desde el preprocesamiento de los datos, el diseño de los modelos, el entrenamiento y validación de los mismos, hasta la evaluación utilizando métricas estándar en clasificación multiclase. Asimismo, se incluyen análisis detallados de errores, matrices de confusión y ejemplos visuales que permiten interpretar mejor el comportamiento de los modelos.

Los resultados muestran que los mejores rendimientos se alcanzan en configuraciones individuales, como fundus individual con DINOv2 (F1-score= 0.79, AUC= 0.94) o fundus con CLIP (F1-score= 0.81, AUC= 0.95), lo que evidencia la solidez de estas modalidades por separado en tareas de clasificación médica. No obstante, algunos modelos de fusión, como la *early fusion* e *intermediate fusion* con DINOv2, han ofrecido resultados competitivos, alcanzando valores de AUC cercanos a 0.90. Estos hallazgos permiten valorar el potencial de la integración multimodal como vía de mejora en determinados escenarios clínicos.

Abstract

The use of medical imaging such as optical coherence tomography (OCT) and fundus photography is essential in the diagnosis and follow-up of ocular diseases. These modalities offer complementary information: while fundus images allow visualization of the retinal surface, OCT provides a cross-sectional view that reveals structural details in depth. However, interpreting these images accurately requires clinical expertise, and analyzing them manually can be error-prone.

The application of artificial intelligence techniques can contribute significantly to the automation of medical diagnosis, providing support to specialists, improving the efficiency of the process and facilitating access to assisted diagnosis in centers with limited resources. In this context, the combination of multimodal information through deep learning algorithms represents a promising line of research.

This Bachelor's Thesis proposes as *baseline* that the combination of fundus and OCT images, by means of feature fusion strategies, allows to improve the performance of the models versus the individual use of each modality. To this end, different fusion techniques (*early*, *intermediate* and *late fusion*) are explored using DINOv2 and CLIP models as feature extractors, trained on a dataset of fundus and OCT images.

The methodology developed ranges from data preprocessing, model design, model training and validation, to evaluation using standard metrics in multiclass classification. Detailed error analysis, confusion matrix and visual examples are also included to better interpret the behavior of the models.

The results show that the best performances are achieved in single configurations, such as single fundus with DINOv2 (F1-score = 0.79, AUC = 0.94) or fundus with CLIP (F1-score = 0.81, AUC = 0.95), evidencing the robustness of these modalities separately in medical classification tasks. However, some fusion models, such as *early fusion* and *intermediate fusion* with DINOv2, have offered competitive results, reaching AUC values close to 0.90. These findings allow us to assess the potential of multimodal integration as an avenue for improvement in certain clinical scenarios.

Capítulo 1

Introducción

La salud visual constituye una prioridad de salud pública a nivel mundial debido a la elevada prevalencia de enfermedades oculares y al impacto que la pérdida de visión tiene en la calidad de vida. Se estima que al menos 2.200 millones de personas en el mundo viven con alguna deficiencia visual o ceguera, de las cuales más de 1.000 millones podrían haberse prevenido o aún no han recibido la atención necesaria [1]. Para atender esta enorme demanda, cada año se realizan millones de exámenes oftalmológicos. Por ejemplo, la tomografía de coherencia óptica (OCT) se ha convertido en un procedimiento estándar en oftalmología y alcanza aproximadamente 30 millones de estudios por año a nivel global [2]. Este volumen masivo de evaluaciones refleja la importancia clínica de la detección y seguimiento oportunos de las patologías oculares, con el fin de prevenir la discapacidad visual evitable.

Entre las enfermedades visuales más prevalentes y de mayor gravedad se encuentran el glaucoma, la degeneración macular asociada a la edad (DMAE), la retinopatía serosa central (RSC) y el edema macular. El glaucoma es la principal causa de ceguera irreversible en el mundo [3]. Estudios epidemiológicos estiman que afectaba a unos 76 millones de personas de 40 a 80 años en 2020, y que esa cifra aumentará a 112 millones para 2040 debido al envejecimiento poblacional [4]. Esta enfermedad, caracterizada por daño progresivo del nervio óptico, suele ser asintomática en fases iniciales, por lo que el diagnóstico temprano resulta crucial para iniciar tratamiento y frenar la progresión antes de que ocurra una pérdida visual permanente [5].

Por su parte, la DMAE es una de las principales causas de deficiencia visual y ceguera en países desarrollados, afectando principalmente a adultos mayores [2]. A nivel global, se calcula que aproximadamente 196 millones de personas padecían DMAE en 2020, y las proyecciones indican un aumento hasta 288 millones en 2040 [6]. La DMAE, especialmente en su forma húmeda o exudativa, puede destruir la visión central de forma rápida, comprometiendo actividades cotidianas básicas; de ahí la necesidad de detectar sus signos iniciales (como las drusas o neovasos coroideos) lo antes posible para implementar terapias intraoculares que preserven la vista.

Otras patologías retinianas, como la retinopatía serosa central (RSC), tienden a presentarse en adultos más jóvenes, a menudo varones en edad laboral. La RSC se caracteriza por el acumulamiento de líquido seroso bajo la retina en la mácula, formando un desprendimiento seroso que distorsiona la visión. Si bien muchos casos de RSC son agudos y autolimitados, un porcentaje no despreciable puede volverse crónico y conducir a daño irreversible. Estudios de seguimiento a largo plazo han revelado que alrededor del 12–13 % de los pacientes con RSC crónica (fluido submacular >6 meses) llegan a desarrollar ceguera legal en ambos ojos [7].

Por último, el edema macular representa una manifestación común a varias enfermedades oculares (por ejemplo, la retinopatía diabética, o uveítis) y consiste en la acumulación de líquido dentro de la retina en la región macular. El edema macular diabético, en particular, es una de las principales causas de pérdida visual en la población en edad productiva de muchos países [8]. Se calcula que en torno al 6–7 % de los pacientes con diabetes desarrollan edema macular diabético, lo que equivale a decenas de millones de personas afectadas en el mundo.

Dado que existen tratamientos eficaces (como terapia láser o fármacos intravítreos antiangiogénicos) que pueden resolver el edema y mejorar la visión, resulta fundamental identificar a estos pacientes lo antes posible. En conjunto, las patologías mencionadas contribuyen significativamente a la carga global de ceguera y discapacidad visual. La detección temprana de todas ellas es esencial para instaurar tratamientos que prevengan o retrasen la pérdida de visión [2].

En este contexto, las imágenes clínicas del fondo de ojo han sido tradicionalmente una herramienta clave para diagnosticar y monitorizar las enfermedades oculares. La fotografía de fundus (retinografía en color) permite visualizar de forma no invasiva la retina, el nervio óptico y la circulación retiniana, identificando hallazgos patológicos como la *excavación del disco óptico* típica del glaucoma (Figura 1.1a), las *drusas y hemorragias* maculares en DMAE (Figura 1.1b), o las *lesiones serosas* en RSC.

(a) Ojo con glaucoma

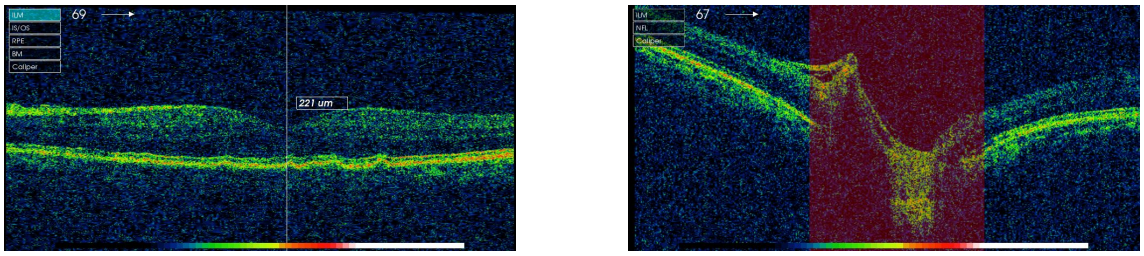
(b) Ojo con DMAE

Figura 1.1: Ejemplos de imágenes fundus utilizadas para el diagnóstico clínico [9]

Del mismo modo, la tomografía de coherencia óptica (OCT) proporciona cortes transversales de alta resolución (del orden de pocos micrómetros) de las capas retinianas y coroideas, revelando cambios sutiles que son invisibles en la fotografía en color [2]. La OCT ha revolucionado el diagnóstico de glaucoma al permitir "disecar" ópticamente tejidos vivos, mostrando, por ejemplo, el *engrosamiento retinal y quistes intrarretinianos* de un edema macular, o el *desprendimiento del epitelio pigmentario* y la presencia de fluido subretiniano en casos de DMAE exudativa [2].

Gracias a su capacidad para detallar la microestructura ocular, la OCT se ha convertido en parte del estándar de cuidado en oftalmología, complementando a la retinografía tradicional. Ambos tipos de imagen (fundus y OCT) se complementan en la detección de patologías: mientras la fotografía de fondo de ojo ofrece una *visión en superficie* útil para el cribado masivo (por ejemplo, en retinopatía diabética), la OCT aporta una *visión seccional* que mejora la sensibilidad diagnóstica, especialmente en etapas iniciales de enfermedades maculares [10], [11]. En suma, la disponibilidad de imágenes digitales de retina de alta calidad ha transformado la forma en que se diagnostican estas enfermedades, permitiendo análisis más detallados y objetivos tanto por parte de especialistas como mediante algoritmos computacionales [12].

Para ilustrar estas aplicaciones, la Figura 3.4 muestra ejemplos de cortes transversales obtenidos mediante OCT, destacando los hallazgos típicos de la DMAE y el glaucoma.



(a) Ojo con DMAE (b) Ojo con glaucoma

Figura 1.2: Ejemplos de cortes transversales obtenidos mediante OCT [9]

1.1. Contexto

Históricamente, la interpretación de imágenes oftalmológicas ha dependido en gran medida de la experiencia y habilidades del especialista en retina, lo que ha generado limitaciones tanto en términos de precisión como de eficiencia diagnóstica. Aunque se han desarrollado herramientas clínicas y protocolos estandarizados para apoyar esta tarea, la lectura de imágenes de fondo de ojo y tomografía de coherencia óptica sigue siendo un proceso complejo, especialmente en estadios tempranos de las enfermedades visuales, donde los signos son sutiles y la variabilidad interobservador puede ser alta. Esta situación ha impulsado en los últimos años el desarrollo de sistemas automatizados basados en inteligencia artificial, y más concretamente, en técnicas de *deep learning*, que han mostrado

resultados muy prometedores en el análisis automático de imágenes de retina.

Uno de los sistemas más representativos es **DeepSeeNet** [13], una red neuronal convolucional desarrollada por investigadores de la Universidad de Iowa y validada con el conjunto de datos AREDS [14], compuesto por más de 58.000 imágenes de fondo de ojo. Esta red fue diseñada específicamente para clasificar los estadios de la DMAE siguiendo el sistema de clasificación de 4 pasos propuesto por el estudio AREDS. En una evaluación de clasificación multiclase basada en pacientes, DeepSeeNet alcanzó una precisión del 67,1 % (kappa=0,558), superando el rendimiento de los especialistas en retina, quienes obtuvieron una precisión del 59,9 % (kappa=0,467) [13].

La Figura 1.3 muestra las matrices de confusión que comparan el rendimiento de DeepSeeNet ajustado y los especialistas en retina en la clasificación de la gravedad de la DMAE.

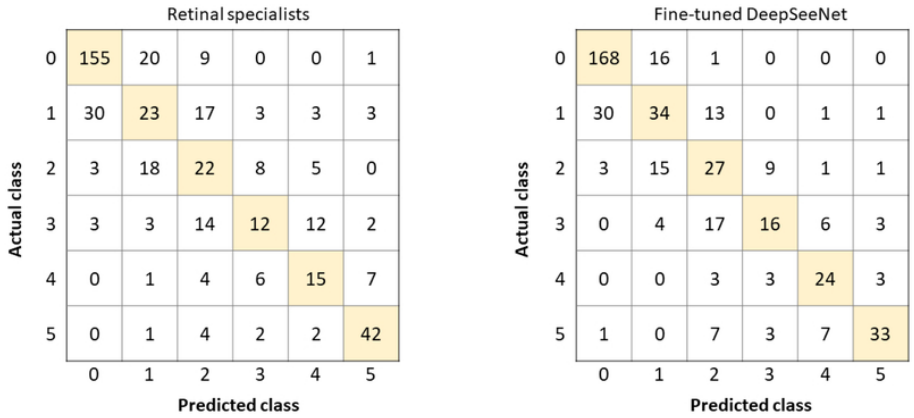


Figura 1.3: Resultados de la comparación entre especialistas en retina y DeepSeeNet. Las columnas y filas de cada matriz son las puntuaciones de escala (0-5) [13]

Por otro lado, competiciones como el desafío *APTOS 2019 Blindness Detection* de la plataforma Kaggle también han impulsado el desarrollo de modelos avanzados para la detección de retinopatía diabética. En esta competición se emplearon modelos como EfficientNet y ResNeXt, entrenados con técnicas de *transfer learning* y estrategias de *data augmentation*, alcanzando puntuaciones de Kappa ponderada superiores a 0,91 [15].

En el ámbito de la tomografía de coherencia óptica, el estudio de **Kermany et al.** [16] marcó un hito importante. En él, se entrenó una red neuronal convolucional con más de 100.000 cortes OCT para clasificar entre retina normal (NORMAL), edema macular diabético (DME), neuritis óptica (CNV) y Drusen (Véase Figura 1.4). El modelo alcanzó una precisión del 96,6 %, con una sensibilidad del 97,8 % y una especificidad del 97,4 %, demostrando el potencial del *deep learning* en tareas de clasificación multiclase complejas.

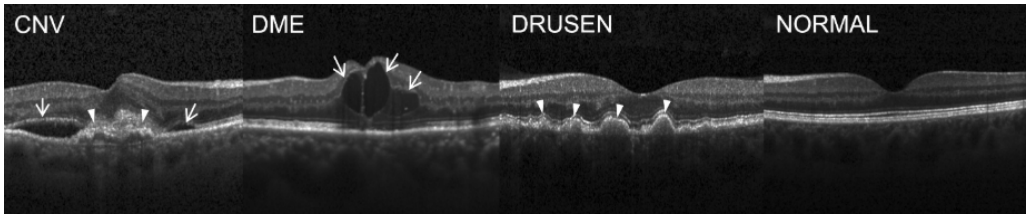


Figura 1.4: Ejemplo de los diferentes tipos de imágenes OCT y sus enfermedades asociadas [16]

Por otro lado, la colaboración entre el hospital Moorfields Eye Hospital y la empresa **DeepMind** resultó en el desarrollo de un sistema de diagnóstico automatizado basado en OCT que, además de detectar la enfermedad, era capaz de recomendar la urgencia de derivación. El modelo, entrenado con más de 15.000 escaneos OCT, logró una concordancia con especialistas superior al 94 %, y fue diseñado con mecanismos de interpretabilidad mediante mapas de atención visual [17], [18].

En estudios más recientes, se ha explorado la clasificación multiclase combinando fotografías de fondo de ojo y escáneres OCT. Por ejemplo, Wang et al. [19] presentaron **MultiEYE**, un sistema que extrae información de imágenes OCT y transfiere ese conocimiento al análisis de fotografías de fondo de ojo sin necesidad de emparejar ambos tipos de datos durante el entrenamiento. Gracias a MultiEYE, la métrica F1 precision-recall aumentó un 4,28 % y la de sensibilidad-especificidad un 3,46 % frente a un modelo basado únicamente en imágenes de fondo de ojo, logrando una F1 media del 62,02 % en la clasificación de nueve enfermedades retinianas. Este tipo de aproximaciones multimodales abren nuevas posibilidades para mejorar la precisión diagnóstica al integrar múltiples fuentes de información clínica.

Después de explorar estas aproximaciones, es importante contextualizar el presente trabajo. Este TFG se apoya en el conjunto de datos proporcionado por Mendeley Data [9], que contiene imágenes fundus y OCT de enfermedades visuales como glaucoma, DMAE, RSC y edema macular. Este conjunto ha sido utilizado en diversos estudios previos y cuenta con anotaciones clínicas expertas.

A diferencia de trabajos anteriores que emplean un único tipo de imagen o modelo, el enfoque adoptado en este proyecto se centra en la fusión de datos fundus y OCT mediante tres estrategias principales: *early fusion*, *intermediate fusion* y *late fusion*. Este tipo de estrategias ha mostrado mejorar los resultados en comparación con el uso de una sola modalidad, ya que la integración de diferentes fuentes de información permite una evaluación más completa y precisa de las patologías oculares, como se ha demostrado en estudios recientes [20]. En la sección 2.5 se explica con mayor detalle este tipo de estrategias.

1.2. Objetivos

El objetivo principal de este Trabajo de Fin de Grado es desarrollar un sistema de detección automática de enfermedades visuales en imágenes de fondo de ojo y OCT. Este objetivo se ha dividido en los objetivos secundarios que se detallan a continuación:

- **Comprender la estructura y contenido de la base de datos de imágenes de fondo de ojo y OCT:** Se ha escogido el conjunto de datos "A Composite Retinal Fundus and OCT Dataset along with Detailed Clinical Markings for Extracting Retinal Layers, Retinal Lesions and Screening Macular and Glaucomatous Disorder" disponible en la plataforma *Mendeley Data* [9].
- **Implementar estrategias de preprocesamiento y normalización de las imágenes de fundus y OCT:** Se procesan las imágenes para facilitar su posterior análisis y el entrenamiento de los modelos.
- **Implementar y ejecutar los métodos de fusión de datos en los modelos:** Se entrenarán los modelos y se aplicarán los diferentes métodos de fusión de datos para producir una predicción más robusta y confiable.
- **Evaluar el rendimiento del sistema de clasificación:** Se valorarán los resultados y se realizarán los correspondientes ajustes necesarios.
- **Analizar cómo los resultados obtenidos se comparan con los de los modelos individuales:** El propósito es comprobar si la combinación de modalidades fundus y OCT, mediante técnicas de fusión multimodal, puede mejorar el rendimiento de los modelos frente al uso aislado de cada tipo de imagen.

Capítulo 2

Marco Teórico

2.1. Detección de patologías en imágenes médicas

El uso de inteligencia artificial, y en particular de técnicas de *deep learning*, ha transformado la forma en la que se analizan imágenes médicas. Hoy en día, los algoritmos son capaces de identificar indicios de enfermedades en diferentes tipos de estudios clínicos, como radiografías de tórax, tomografías, imágenes *fundus* del fondo de ojo u OCT, entre otros. Gracias a estas técnicas, se pueden detectar patrones visuales característicos de patologías como neumonía, fracturas, tumores o retinopatía diabética. Esto permite no solo automatizar parte del diagnóstico, sino también aplicarlo a gran escala en programas de cribado y reducir la carga de trabajo del personal sanitario, sin comprometer la calidad del diagnóstico [13], [21]. Modelos como las redes neuronales convolucionales (CNN) o los transformers se han consolidado como herramientas eficaces para este tipo de tareas, permitiendo incluso alcanzar resultados comparables a los de especialistas humanos en algunos contextos clínicos [22], [23]. Como ejemplo de la detección automática de patologías mediante inteligencia artificial, la Figura 2.1 muestra un caso de identificación de enfermedades en radiografías de tórax.

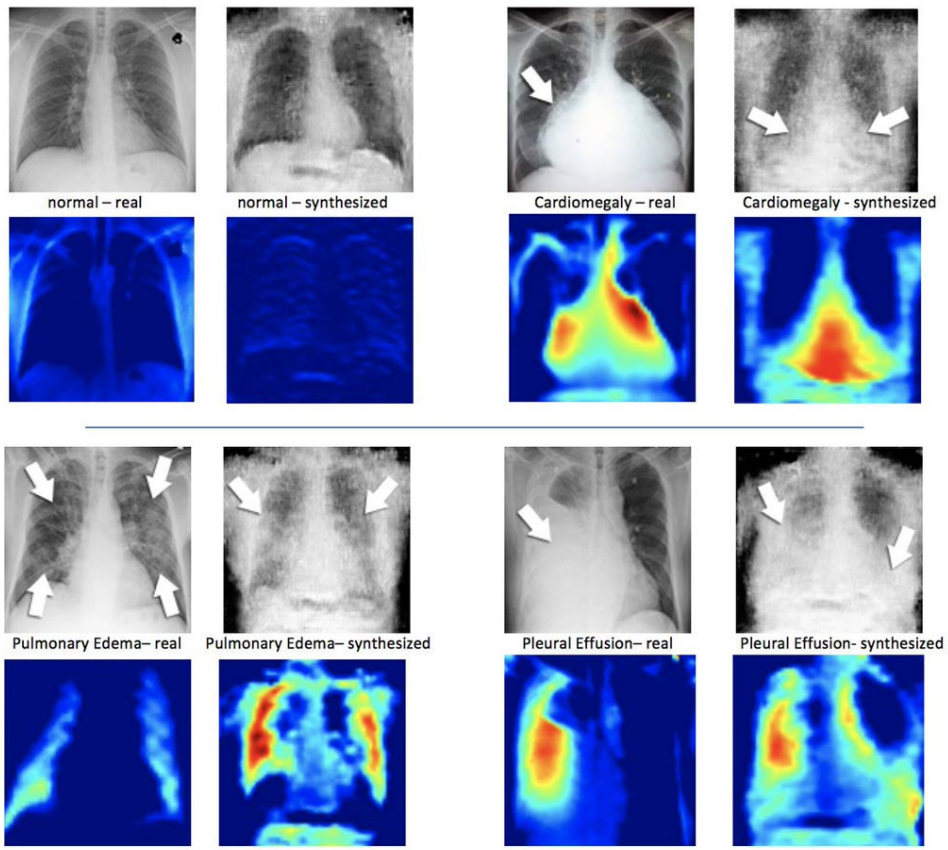


Figura 2.1: Ejemplo de detección automática de patologías en radiografía de tórax [22].

A nivel técnico, existen distintas formas de abordar la detección automática de patologías. Una opción es entrenar un modelo supervisado desde cero utilizando grandes conjuntos de imágenes ya etiquetadas. Otra posibilidad muy extendida consiste en reutilizar modelos preentrenados y afinarlos con imágenes específicas de la tarea médica, lo que reduce la necesidad de datos y tiempo de entrenamiento. También se emplean técnicas no supervisadas que buscan anomalías en la imagen sin necesidad de anotaciones, o enfoques auto-supervisados y de contraste que extraen representaciones útiles sin necesidad de intervención humana [24], [25]. Por último, se están desarrollando modelos multimodales que combinan distintas modalidades de imagen o, incluso, incluyen información textual y clínica del paciente para enriquecer el análisis [26].

2.2. Redes Neuronales Convolucionales

Las redes neuronales convolucionales, conocidas como CNN (*Convolutional Neural Networks*), constituyen una arquitectura diseñada específicamente para procesar información estructurada como matrices multidimensionales, destacando particularmente en la tarea de análisis de imágenes. Una CNN está compuesta generalmente por tres capas esenciales [27]:

1. **Capa convolucional**: Es el núcleo principal de la CNN. Realiza convoluciones sobre los datos de entrada con un filtro (o *kernel*), generando mapas de características mediante operaciones que extraen rasgos significativos como bordes, texturas o formas. Tras esta operación, suele aplicarse una función de activación no lineal (por ejemplo, ReLU), facilitando la extracción de representaciones más complejas mediante la combinación de múltiples capas convolucionales.
2. **Capa de *pooling***: Esta capa tiene como objetivo reducir las dimensiones espaciales de los mapas de características, utilizando operaciones estadísticas como máximo (*max pooling*) o promedio (*average pooling*). Esto contribuye a disminuir la complejidad computacional y el volumen de datos, conservando la información más relevante.
3. **Capa completamente conectada**: En esta etapa final, cada neurona se conecta directamente a todas las neuronas de la capa anterior. Es aquí donde se efectúa la clasificación mediante la combinación de todas las características extraídas anteriormente. Habitualmente, utiliza una función de activación como *softmax*, generando así probabilidades asociadas a las diferentes categorías posibles.

Estas tres capas se pueden apreciar en el diagrama simplificado de la arquitectura de una red convolucional de la Figura 2.2

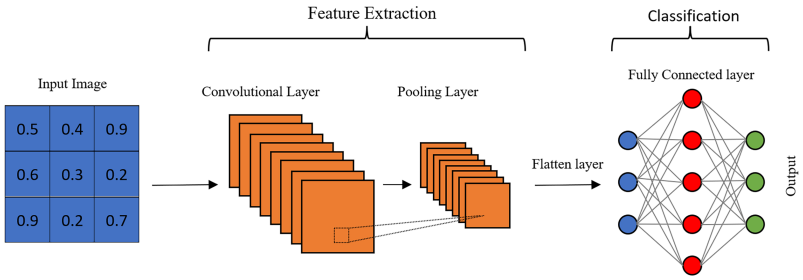


Figura 2.2: Ejemplo de arquitectura de una red convolucional [28]

2.2.1. Operación de Convolución

La convolución es una operación esencial en redes neuronales convolucionales (CNN), que permite extraer características locales de una imagen mediante la aplicación de filtros (también llamados kernels). Cada filtro recorre la imagen y genera un mapa de activación que resalta ciertas propiedades visuales como bordes, texturas o patrones repetitivos. Matemáticamente, se puede expresar como una suma ponderada de los valores de píxeles en la vecindad de cada punto, multiplicados por los coeficientes del filtro aplicado (Figura 2.3) [29].

Durante el entrenamiento, los valores de los filtros se ajustan automáticamente para maximizar el rendimiento del modelo en la tarea específica, lo que permite detectar patrones cada vez más complejos en las capas más profundas. Esta operación es especialmente poderosa porque explota la estructura espacial de las imágenes y reduce drásticamente el número de parámetros en comparación

con redes densas. La convolución también favorece la reutilización de parámetros y la invariancia local, lo que la hace ideal para problemas de clasificación y segmentación en visión por computador [30], [31].

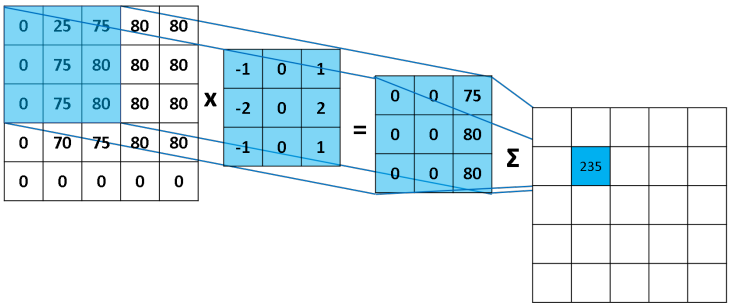


Figura 2.3: Operación de convolución [32]

2.2.2. Operación *pooling*

La operación *pooling* es una técnica que busca reducir la dimensión de los mapas de características, manteniendo lo más relevante y descartando detalles menos importantes. Generalmente, se implementa tomando el valor máximo (*max pooling*) o el promedio (*average pooling*) de las regiones adyacentes del mapa de características, lo que simplifica la información procesada por la red. Aunque esta reducción facilita el procesamiento, puede también eliminar ciertos detalles finos de la información original.

2.3. Transformers

Los modelos *transformers* representan una arquitectura de redes neuronales inicialmente diseñada para el procesamiento de lenguaje natural, aunque su aplicabilidad se ha extendido exitosamente a otras áreas como la visión por computador. En particular, los *Visual Transformers* se han desarrollado específicamente para tareas que implican datos visuales [33].

Una de las características distintivas de los *transformers* es su capacidad para capturar relaciones a largo plazo entre los elementos de una secuencia, permitiendo además un procesamiento paralelo eficiente. Esta propiedad se debe al uso del mecanismo de *autoatención*, el cual evalúa la relevancia de cada elemento en relación con los demás dentro de una secuencia. A través de esta técnica, es posible codificar de manera contextual cada componente, integrando tanto información local como global [34].

El mecanismo de atención utilizado se conoce como *Scaled Dot-Product Attention*, y se define mediante la siguiente expresión [35]:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{2.1}$$

donde Q, K y V representan respectivamente las matrices de consultas (*queries*), claves (*keys*) y valores (*values*), cada una derivada de la entrada original a través de proyecciones lineales. El término d_k corresponde a la dimensión de las claves, y se utiliza para escalar la similitud entre Q y K, evitando que los valores crezcan demasiado y dificulten el aprendizaje. La operación de *softmax* convierte estas similitudes en una distribución de probabilidad, que pondera la importancia relativa de cada valor V al generar la salida [35].

Gracias a este mecanismo, el modelo puede centrarse dinámicamente en las partes más relevantes de la entrada, lo que ha demostrado ser muy útil en tareas donde las relaciones entre elementos no son locales, como en visión por computador [35].

A diferencia de las convoluciones tradicionales, donde los filtros son estáticos, en los *transformers* los filtros se generan dinámicamente en función de la entrada. Esta flexibilidad ha demostrado ser especialmente útil en tareas como la detección de objetos, donde se tratan las relaciones entre distintos elementos como un problema de predicción de conjuntos [36].

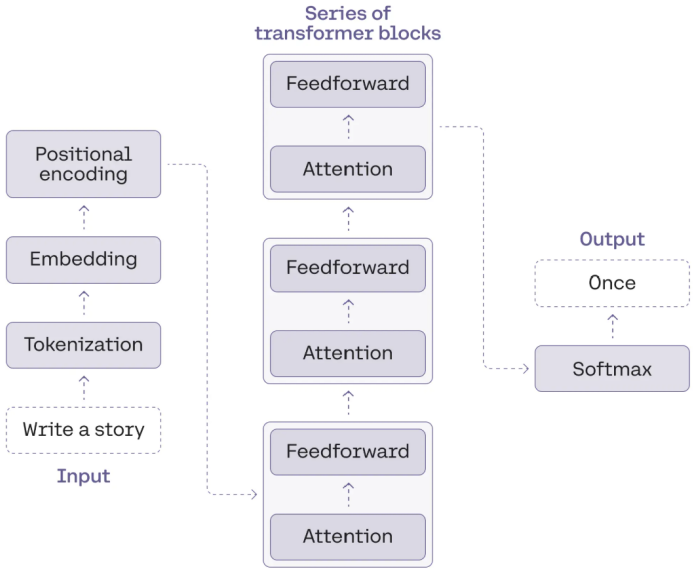


Figura 2.4: Arquitectura simplificada de un modelo *transformer* [37].

La Figura 2.4 muestra una arquitectura simplificada de un modelo *transformer*, en la cual los datos de entrada son primero tokenizados y luego convertidos en vectores mediante un proceso de *embedding*. Posteriormente, se añade codificación posicional para conservar la información secuencial, y finalmente, la información se procesa a través de bloques compuestos por capas de atención y redes completamente conectadas.

2.3.1. Vision Transformers (ViT)

Los modelos *Vision Transformers* (ViT) constituyen una adaptación de la arquitectura *transformer*, originalmente concebida para el procesamiento de lenguaje natural, al dominio de la visión por computador [38]. Esta adaptación ha permitido aprovechar la capacidad de los *transformers* para capturar relaciones globales en secuencias, aplicándola al análisis de imágenes mediante un esquema novedoso de representación y procesamiento [38].

Una de las características principales de los ViT es que eliminan el uso de convoluciones, típico de las redes neuronales convolucionales (CNN), y dividen las imágenes en bloques de tamaño fijo denominados *patches* [39]. Cada *patch* es aplanado y proyectado linealmente en un espacio de características. Posteriormente, a estos vectores se les añade una codificación posicional que preserva la información espacial de la imagen. Finalmente, la secuencia resultante se procesa mediante una arquitectura basada en bloques de atención y redes completamente conectadas, siguiendo el mismo principio de diseño que en los *transformers* tradicionales [39].

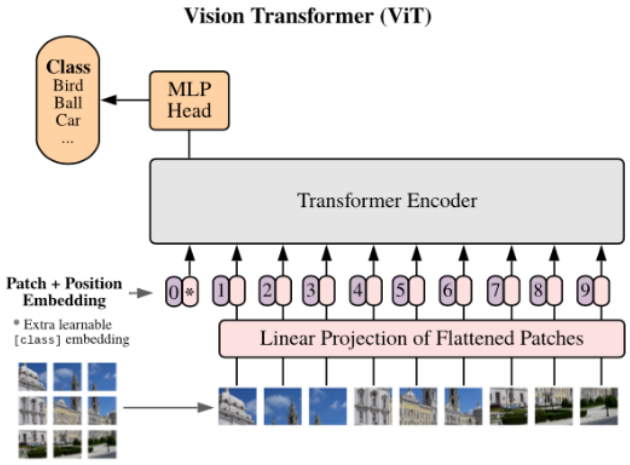


Figura 2.5: Esquema simplificado de la arquitectura Vision Transformer (ViT) [39].

La Figura 2.5 muestra una arquitectura simplificada del ViT, en la cual los datos de entrada, en este caso, una imagen, se segmentan en *patches* y son transformados en vectores mediante un proceso de proyección lineal. Posteriormente, se incorpora una codificación posicional que permite conservar la estructura espacial de los datos. Esta representación secuencial se introduce en una red formada por bloques sucesivos de atención y capas *feedforward*, al final de los cuales se extrae una representación global que puede emplearse para tareas de clasificación, segmentación u otras aplicaciones clínicas. Esta arquitectura ha demostrado ser especialmente eficaz en contextos donde resulta crucial capturar relaciones espaciales de largo alcance, como es el caso del análisis de imágenes médicas [38].

2.4. Modelos utilizados en este trabajo

En los últimos años, han surgido modelos fundamentales en *computer vision* que aprenden representaciones genéricas a partir de grandes conjuntos de datos. A continuación, se describen dos ejemplos destacados de estos modelos: DINOv2 (aprendizaje autosupervisado puramente visual) y CLIP (aprendizaje multimodal contrastivo texto-imagen).

2.4.1. DINOv2 (Self-Distillation with No Labels)

DINOv2 (Distillation with no labels) es un modelo de aprendizaje auto-supervisado desarrollado por Meta AI para tareas de visión por computador. Su principal característica es que aprende directamente desde imágenes no etiquetadas, utilizando un esquema de auto-distilación tipo *profesor-estudiante*. Esta técnica permite que una red (el estudiante) aprenda a imitar las representaciones generadas por otra red (el profesor), sin necesidad de etiquetas humanas [40], [41].

La arquitectura base de DINOv2 se fundamenta en los Vision Transformers (ViT). Tanto el estudiante como el profesor comparten esta arquitectura ViT, pero sus parámetros difieren: el profesor es una versión suavizada del estudiante, actualizada mediante promedio móvil exponencial (EMA) [42]. DINOv2 incorpora además un doble objetivo de entrenamiento: uno a nivel de imagen (global) y otro a nivel de parches (local), lo que permite capturar tanto la semántica general como los detalles finos de una imagen [43].

El funcionamiento del modelo DINOv2 puede observarse de forma simplificada en el diagrama de la Figura 2.6.

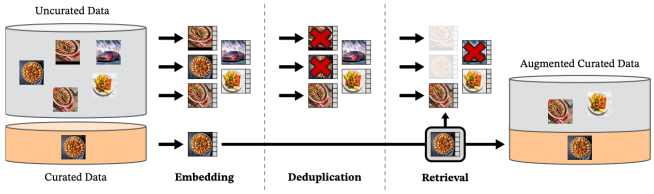


Figura 2.6: Esquema de funcionamiento del modelo DinoV2 [44].

El modelo fue entrenado sobre un conjunto curado de 142 millones de imágenes no etiquetadas, logrando resultados de estado del arte en múltiples tareas. Entre sus aplicaciones destacan la clasificación de imágenes mediante k-NN o clasificadores lineales, la segmentación semántica sin etiquetas, la detección de objetos y la estimación de profundidad monocular. Gracias a la calidad de sus embeddings, DINOv2 puede utilizarse directamente, sin necesidad de reentrenamiento, como extractor de características en sistemas visuales avanzados [40].

DINOv2 representa una evolución respecto a DINO original, incorporando mejoras como el objetivo local (inspirado en iBOT [45]), técnicas de normalización avanzadas como Sinkhorn-Knopp [46], enmascaramiento de parches, curriculum de resolución [47] y una familia de modelos escalables que van desde la ver-

sión pequeña (ViT-S) hasta la versión grande (ViT-g) de los Vision Transformers. Estas características lo convierten en uno de los modelos auto-supervisados más eficientes y versátiles del panorama actual [41].

2.4.2. CLIP (Contrastive Language–Image Pretraining)

CLIP (Contrastive Language–Image Pretraining) es un modelo multimodal desarrollado por OpenAI que permite asociar imágenes con descripciones en lenguaje natural. A diferencia de los clasificadores tradicionales que aprenden a partir de etiquetas específicas para cada clase, CLIP se entrena con pares imagen-texto obtenidos de internet, sin necesidad de anotaciones estructuradas. Su objetivo es aprender representaciones visuales generalistas que puedan transferirse fácilmente a distintas tareas visuales mediante el uso del lenguaje como supervisión [48], [49].

La arquitectura de CLIP está formada por dos componentes principales: un *encoder* de imágenes y un *encoder* de texto. El *encoder* de imágenes puede basarse en una arquitectura ResNet o en un Vision Transformer (ViT), mientras que el *encoder* textual suele emplear un transformer similar a los utilizados en modelos de lenguaje como GPT. Cada entrada (imagen o texto) se procesa de forma independiente y se proyecta en un espacio vectorial compartido de dimensiones fijas. Durante el entrenamiento, CLIP utiliza un objetivo contrastivo: se calcula la similitud entre todos los pares imagen-texto en un lote, y se optimiza para que los pares correctos (imagen y texto que se corresponden) tengan una mayor similitud que los pares no coincidentes [48].

El resultado es un modelo capaz de entender el contenido semántico de una imagen en relación con descripciones en lenguaje natural. Por ejemplo, al recibir una imagen y varias frases como “un perro”, “una bicicleta”, “una persona corriendo”, CLIP calcula qué frase se aproxima más a la imagen en el espacio de embeddings, y devuelve la más adecuada sin necesidad de haber sido entrenado explícitamente en esas clases (Figura 2.7). Esta capacidad se conoce como *zero-shot classification* y permite aplicar el modelo a nuevas tareas sin reentrenamiento [49].

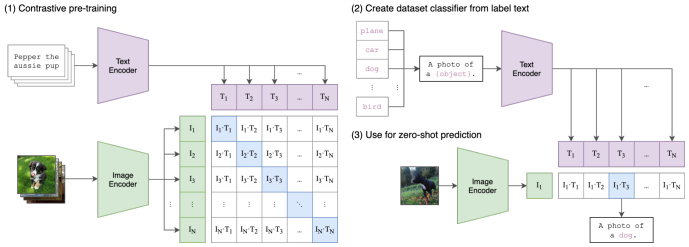


Figura 2.7: Esquema de funcionamiento del modelo CLIP [50].

Además de la clasificación en *zero-shot*, CLIP puede emplearse en tareas de búsqueda semántica, emparejamiento imagen-texto, recuperación de imágenes a partir de texto y etiquetado automático. Por ejemplo, en una base de datos de imágenes, se puede escribir una consulta como “una persona caminando bajo

la lluvia", y CLIP devolverá las imágenes más similares a esa descripción. De igual modo, también permite generar descripciones asociadas a una imagen mediante selección entre frases candidatas. Esta versatilidad ha hecho que CLIP se utilice como base en numerosos sistemas multimodales actuales y en modelos generativos que combinan visión y lenguaje [48].

2.5. Estrategias de fusión de datos en visión por computador

En visión por computador y campos relacionados, es habitual disponer de múltiples fuentes o modalidades de datos, como imágenes provenientes de diferentes cámaras o información textual asociada. Para trabajar eficazmente esta información multimodal, se emplean diferentes estrategias de fusión de datos, las cuales se distinguen por el punto específico dentro del flujo de procesamiento del modelo en el que se combinan los datos. Las tres estrategias principales son: *early fusion*, *intermediate fusion*, y *late fusion* [51]-[53].

2.5.1. Early Fusion

La *early fusion* combina las diversas modalidades en las etapas iniciales del procesamiento antes del aprendizaje profundo del modelo [52], [53]. En términos prácticos, esta estrategia consiste en fusionar las modalidades directamente en la etapa de entrada o en capas iniciales del modelo (Figura 2.8). Por ejemplo, imágenes RGB junto con mapas de profundidad pueden concatenarse como canales adicionales de entrada y luego alimentar una red neuronal única con esta representación conjunta.

La principal ventaja radica en que permite aprender desde las primeras capas las interacciones entre modalidades, aprovechando correlaciones inmediatas en los datos. Sin embargo, también conlleva retos significativos, como la necesidad de un riguroso preprocesamiento y sincronización entre fuentes para asegurar compatibilidad espacial y temporal [52]. Además, si las modalidades tienen diferentes escalas o frecuencias, combinarlas tempranamente puede ser complicado y susceptible al ruido.

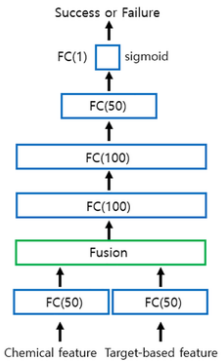


Figura 2.8: Representación gráfica de la fusión de datos *early fusion* [54].

2.5.2. Intermediate Fusion

La *intermediate fusion* procesa cada modalidad por separado inicialmente, hasta generar representaciones latentes de alto nivel, que posteriormente se combinan en alguna capa oculta intermedia del modelo [51]. Es una técnica comúnmente utilizada en modelos profundos multimodales, donde cada modalidad cuenta con subredes independientes que extraen características específicas (por ejemplo, CNN para imágenes y LSTM para texto).

Posteriormente, estas características se combinan en *fusion layers* o capas compartidas, permitiendo un procesamiento conjunto a partir de ahí (Figura 2.9). La ventaja más destacada de esta técnica es su flexibilidad, pues permite experimentar con distintos niveles y profundidades de fusión según la tarea y características de los datos. No obstante, presenta una mayor complejidad arquitectónica y puede ser susceptible al sobreajuste debido a la alta dimensionalidad de las representaciones combinadas [51]. A pesar de esto, la *intermediate fusion* ha demostrado ser eficaz en tareas donde la interacción entre modalidades es crucial, generando representaciones más ricas y profundas comparadas con las estrategias de *early* o *late fusion*.

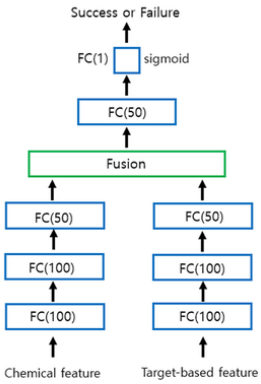


Figura 2.9: Representación gráfica de la fusión de datos *intermediate fusion* [54].

2.5.3. Late Fusion

La *late fusion* integra las modalidades en etapas avanzadas del procesamiento, típicamente en la fase final de toma de decisiones o en la etapa de predicción del modelo [51]. En este enfoque, cada modalidad se procesa individualmente mediante modelos especializados, generando salidas o decisiones independientes que luego se combinan (Figura 4.2). La combinación final puede realizarse mediante diversas técnicas, como promedios ponderados, votaciones o incluso mediante un clasificador adicional.

Inspirada en métodos de *ensemble*, esta estrategia es sencilla, modular y robusta, permitiendo que cada modelo especializado opere independientemente. Esta independencia significa que errores en una modalidad no afectan directamente a otras, proporcionando considerable tolerancia a fallos [52]. Además, la *late fusion* es altamente flexible en integrar fuentes heterogéneas, al no requerir es-

tricta sincronización de formatos o escalas. Una desventaja, sin embargo, es que no aprovecha interacciones tempranas entre modalidades y, en algunos casos, el rendimiento puede no superar al del mejor modelo individual si la combinación no se ejecuta adecuadamente. No obstante, la *late fusion* es ampliamente utilizada cuando se requiere simplicidad en la implementación, clara separación de responsabilidades o cuando existen incompatibilidades en los ritmos y formatos de las modalidades que hacen inviable la *early fusion*.

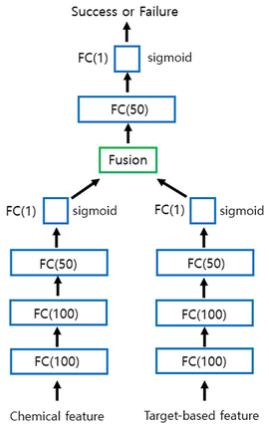


Figura 2.10: Representación gráfica de la fusión de datos *late fusion* [54].

2.6. Métricas de evaluación

2.6.1. Precisión

La precisión evalúa la frecuencia con la que las predicciones positivas realizadas por el modelo son correctas. Su cálculo se realiza dividiendo el número de verdaderos positivos entre el total de casos clasificados como positivos, incluyendo los falsos positivos:

$$Precision = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Positivos} \tag{2.2}$$

El valor de la *precisión* varía entre 0 y 1 o como porcentaje, siendo preferible un valor más alto.

2.6.2. Recall

El *recall* es una métrica que mide la capacidad del modelo para identificar correctamente los verdaderos positivos, es decir, los casos realmente positivos en el conjunto de datos. Para su cálculo, se divide la cantidad de verdaderos positivos entre la suma de verdaderos positivos y falsos negativos, estos últimos siendo casos positivos que no fueron detectados:

$$Recall = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Negativos} \tag{2.3}$$

Al igual que con la *precisión*, su valor varía entre 0 y 1 o puede expresarse como un porcentaje, siendo preferible un valor elevado.

2.6.3. F1-Score

El *F1-Score* representa la media armónica entre precisión y *recall*, combinando ambas métricas en un único valor equilibrado. Su valor óptimo es 1 y su peor valor es 0. Se calcula mediante la siguiente fórmula:

$$F1\text{-}Score = 2 \cdot \frac{Precision \times Recall}{Precision + Recall} \tag{2.4}$$

2.6.4. Área bajo la curva ROC

El *Área Bajo la Curva ROC* es una métrica que indica qué tan bien un modelo puede distinguir entre clases positivas y negativas. Representa la probabilidad de que una instancia positiva reciba una puntuación mayor que una negativa seleccionada al azar, calculado de la siguiente forma:

$$AUC = \int_0^1 TPR(FPR^{-1}(x))\,dx \tag{2.5}$$

Su valor se encuentra entre 0 y 1, siendo preferible que sea cercano a 1, ya que indica una mayor capacidad del modelo para discriminar correctamente entre las clases, al igual que con la *precisión* y el *recall*.

Capítulo 3

Patologías Oculares en OCT y Fundus

3.1. Enfermedades oculares a detectar en las imágenes OCT y Fundus

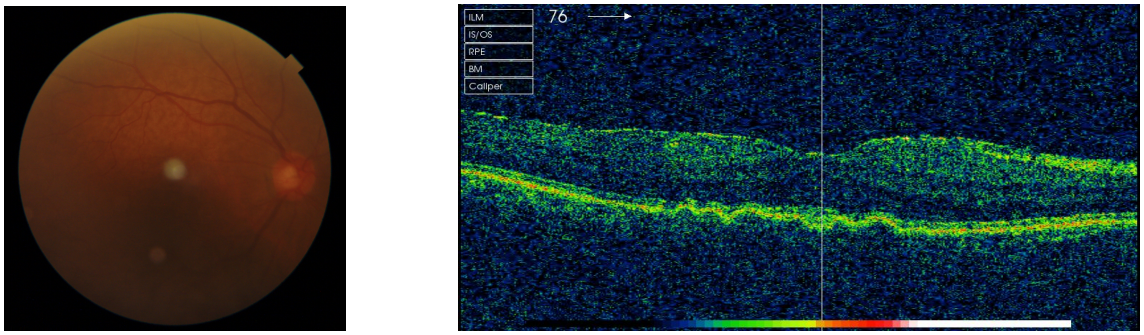
La tomografía de coherencia óptica (OCT) y la fotografía del fondo de ojo son herramientas fundamentales para diagnosticar diferentes patologías retinianas. A continuación, se presentan las cuatro enfermedades a detectar en este trabajo.

3.1.1. Degeneración macular asociada a la edad (DMAE)

La degeneración macular asociada a la edad (DMAE) es una enfermedad ocular degenerativa que afecta la mácula, la zona central de la retina responsable de la visión detallada y central. Debido al daño progresivo de las células en la mácula, la DMAE provoca una pérdida de agudeza en la visión central donde los pacientes pueden notar una mancha borrosa en el centro de su campo visual o que las líneas rectas aparecen onduladas. Es una patología relacionada con el envejecimiento, muy común en personas mayores (generalmente a partir de los 50-60 años) [55]. De hecho, la DMAE constituye una de las principales causas de pérdida de visión entre los adultos mayores en los países desarrollados [56]. Cabe destacar que esta enfermedad no produce dolor y no afecta a la visión periférica; sin embargo, la disminución de la visión central dificulta actividades cotidianas como la lectura, conducir o reconocer caras [56].

Existen dos formas principales de DMAE: la seca (atrófica) y la húmeda (exudativa) [57]. La DMAE seca representa cerca del 80 %-90 % de los casos. Se caracteriza por una atrofia gradual de las células retinianas de la mácula y la acumulación de drusas. Su curso suele ser lento y progresivo [56]. En cambio, la DMAE húmeda es menos frecuente pero más agresiva; en ella se forman vasos sanguíneos anormales debajo de la retina (neovasos coroideos) que gotean sangre y líquido debajo de la mácula [57]. Esto provoca un deterioro rápido de la visión central y, si no se trata, puede generar cicatrices que destruyen la función

macular [56]. Los síntomas de la forma húmeda suelen incluir una distorsión notable de las imágenes (metamorfopsia) y una pérdida súbita de visión central más pronunciada. Actualmente no existe una cura que recupere la visión perdida, pero sí tratamientos, como vitaminas antioxidantes en la forma seca o inyecciones intraoculares antiangiogénicas en la forma húmeda, que pueden retrasar la progresión de la enfermedad y mejorar el pronóstico visual [56], [57].



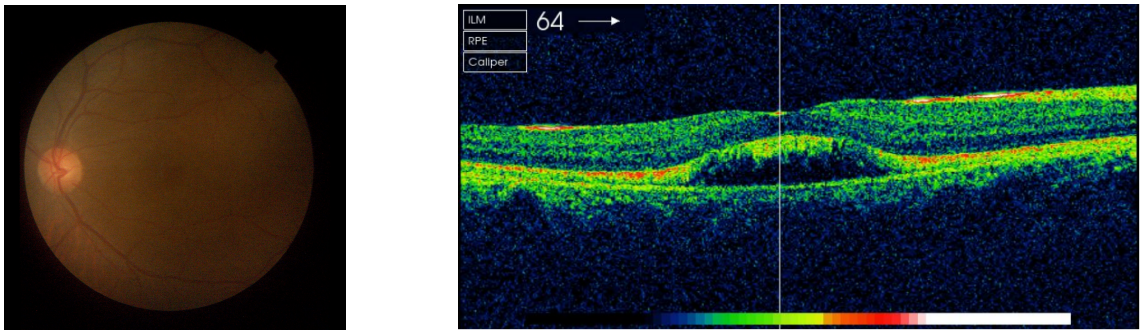
(a) Fundus con DMAE (b) OCT con DMAE

Figura 3.1: Ejemplos de imagenes fundus y OCT con DMAE [9]

3.1.2. Coriorretinopatía serosa central (CSR)

La coriorretinopatía serosa central (CSR) es una enfermedad de retina no hereditaria que suele afectar a adultos jóvenes y de mediana edad, con mayor frecuencia en varones entre los 20 y 50 años de edad [58]. En esta enfermedad se acumula líquido seroso debajo de la retina neurosensorial, específicamente bajo la mácula, produciendo un desprendimiento localizado de la misma [58]. Como resultado, la visión central del ojo afectado se ve disminuida: el paciente típicamente percibe una mancha borrosa u oscura en el centro de su campo visual (escotoma central) y distorsión de las imágenes (las líneas rectas pueden verse torcidas, fenómeno llamado metamorfopsia). La CSR suele presentarse de forma unilateral (un solo ojo) y, en sus fases agudas, no suele causar dolor ni enrojecimiento ocular, solamente alteraciones visuales. A pesar de la merma en la agudeza visual central, en muchos casos, la visión puede recuperarse parcialmente una vez que la lesión se resuelve.

La causa exacta de la CSR no está completamente clara, pero se ha asociado con situaciones de estrés elevado y prolongado. Se postula que una sobreproducción de cortisol y otras hormonas del estrés provoca cambios en la coroides (capa vascular bajo la retina) que llevan a fugas de fluido bajo la retina [58]. De igual manera, el uso de medicamentos como la cortisona se ha identificado como un factor desencadenante en personas predispuestas [58]. En la mayoría de los pacientes, la CSR es una condición autolimitada: el desprendimiento seroso suele reabsorberse espontáneamente y la visión central se recupera en gran parte dentro de algunas semanas o meses. De hecho, en alrededor del 90 % de los casos, la mejoría ocurre sin necesidad de tratamiento en un plazo de 2 a 3 meses [59]. Sin embargo, la enfermedad puede reaparecer, derivando en que aproximadamente la mitad de los pacientes sufren recurrencias en el futuro [60].



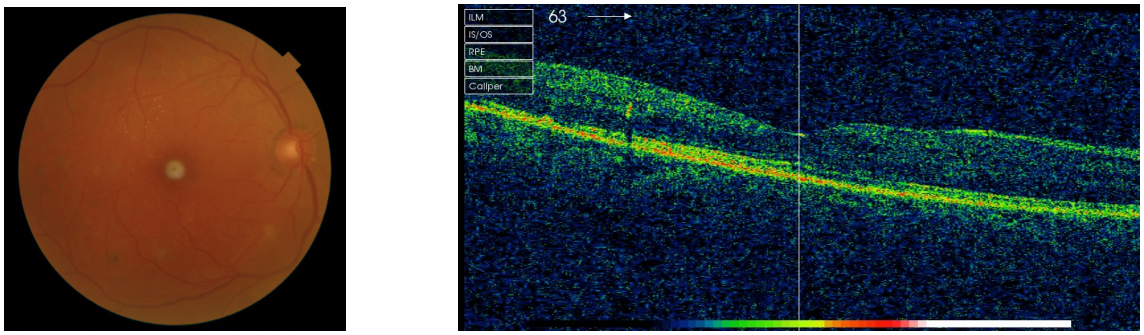
(a) Fundus con CSR (b) OCT con CSR

Figura 3.2: Ejemplos de imagenes fundus y OCT con CSR [9]

3.1.3. Edema macular (ME)

Se denomina edema macular a la acumulación anormal de fluido en la mácula. Esta acumulación de líquido ocurre cuando los capilares retinianos de la mácula filtran suero debido a una alteración en su permeabilidad, provocando hinchazón (edema) en dicha región [61]. Como resultado, la retina se engrosa en el área macular y la visión central se vuelve borrosa. Los pacientes con edema macular típicamente experimentan disminución de la agudeza visual central, distorsión de las formas (las líneas rectas pueden percibirse onduladas) y dificultades para leer o reconocer detalles, e incluso pueden notar los colores apagados. Característicamente, estos síntomas no van acompañados de dolor ocular, dado que el edema macular en sí mismo no causa sensibilidad ni rojez [61].

El edema macular no es una enfermedad independiente, sino que suele aparecer como complicación de diversas patologías oculares. La causa más frecuente es la retinopatía diabética: de hecho, el edema macular diabético es la principal causa por la que las personas con diabetes pierden visión [62]. Otros cuadros asociados con edema en la mácula incluyen las trombosis u oclusiones de las venas retinianas (como la oclusión de vena central de la retina), las uveítis crónicas (inflamaciones intraoculares), la cirugía ocular reciente (por ejemplo, puede ocurrir edema macular tras una operación de cataratas, conocido como síndrome de Irvine-Gass), las distrofias hereditarias de la retina como la retinosis pigmentaria y la propia degeneración macular asociada a la edad en etapas avanzadas [61].



(a) Fundus con ME (b) OCT con ME

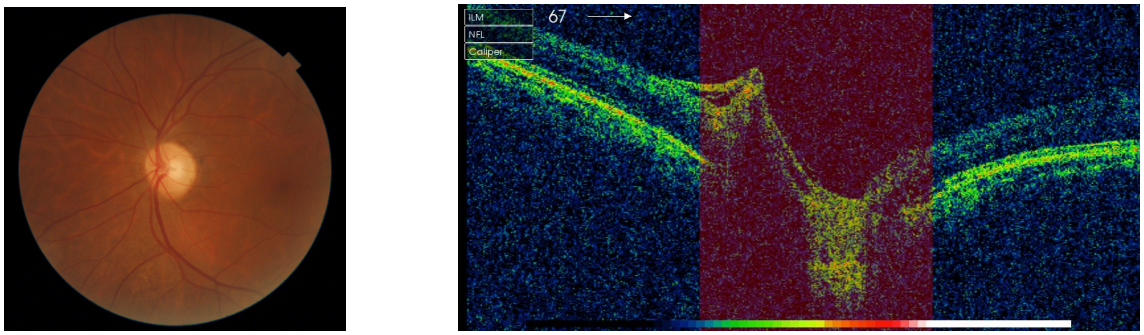
Figura 3.3: Ejemplos de imagenes fundus y OCT con ME [9]

3.1.4. Glaucoma

El glaucoma describe un grupo de enfermedades oculares que tienen en común el daño progresivo del nervio óptico, generalmente asociado a una presión intraocular elevada. En condiciones normales, el ojo produce un líquido transparente (humor acuoso) que drena continuamente; si este drenaje se dificulta, aumenta la presión dentro del globo ocular y puede lesionar las fibras del nervio óptico. Con el tiempo, ese daño se traduce en una pérdida irreversible de visión. El glaucoma típicamente afecta primero la visión periférica (campimetría): los pacientes pueden no notar nada al inicio, pero gradualmente desarrollan reducción del campo visual lateral, avanzando luego hacia la visión central en etapas tardías. Es una de las principales causas de ceguera irreversible a nivel mundial [63]. Un aspecto engañoso del glaucoma es su falta de síntomas iniciales: en el glaucoma de ángulo abierto (la forma más común), la elevación de la presión ocular ocurre de forma silenciosa, sin dolor ni molestias, de modo que el paciente solo nota la pérdida de visión cuando esta ya es avanzada [63], [64]. Por ello, al glaucoma se le llama a veces el “ladrón silencioso de la vista”.

Se reconocen varios tipos de glaucoma. El más frecuente es el glaucoma primario de ángulo abierto, responsable de la mayoría de los casos, especialmente en adultos mayores [65]. En este tipo, el ángulo de drenaje del ojo (entre la córnea y el iris) está abierto pero el flujo del humor acuoso se dificulta por cambios microscópicos, generando un incremento lento pero sostenido de la presión intraocular. El glaucoma de ángulo abierto suele ser bilateral y tiene un componente genético: el riesgo es mayor en personas con antecedentes familiares de glaucoma, y ciertas poblaciones como los individuos de ascendencia africana presentan predisposición más alta a desarrollarlo [66]. Otro subtipo es el glaucoma de ángulo cerrado, en el cual el ángulo de drenaje ocular se bloquea de forma abrupta. Esto provoca un aumento repentino de la presión intraocular y se manifiesta con síntomas agudos: dolor ocular intenso, ojo rojo, visión borrosa con halos coloreados alrededor de las luces, náuseas y vómitos. El glaucoma de ángulo cerrado constituye una emergencia médica que requiere tratamiento inmediato para evitar daños permanentes [64], [66]. Además, existen glaucomas secundarios, causados por factores identificables como traumatismos oculares, uso prolongado de corticoides, inflamaciones intraoculares (uveítis) o complica-

ciones vasculares (por ejemplo, neovascularización en la diabetes) [66]. También puede presentarse glaucoma congénito en bebés, debido a malformaciones en el desarrollo del ojo; aunque es raro, debe sospecharse si un recién nacido tiene opacidad corneal, aumento del tamaño ocular o sensibilidad a la luz en los primeros meses de vida [66]. Dado que actualmente no existe cura para el glaucoma (la pérdida de visión ya instaurada no se puede recuperar), el objetivo del tratamiento es prevenir o ralentizar el daño adicional. Esto se logra controlando la presión intraocular con colirios, láser o cirugía según el caso [63].



(a) Fundus con glaucoma (b) OCT con glaucoma

Figura 3.4: Ejemplos de detección de Glaucoma [9]

3.1.5. Interpretación de las imágenes fundus y OCT

La imagen fundus, o retinografía, es una fotografía en color del fondo del ojo. Muestra una vista general de la retina desde el frente, lo que permite observar estructuras importantes como el nervio óptico, los vasos sanguíneos, la mácula y el resto del tejido retiniano [67]. El nervio óptico se ve como un círculo de color claro situado hacia un lado de la imagen (zona nasal). Su forma y tamaño pueden dar pistas sobre enfermedades como el glaucoma, especialmente si se observa una excavación central agrandada [65]. La mácula, encargada de la visión central, está situada más al centro y tiene una tonalidad más oscura. Los vasos retinianos parten del nervio óptico y se extienden por toda la retina.

En esta imagen se pueden detectar distintos signos que ayudan a identificar enfermedades:

- **Drusas**: pequeños puntos amarillos en la zona central del ojo (mácula), que pueden indicar degeneración macular asociada a la edad (DMAE) [56].
- **Exudados**: manchas blancas bien definidas causadas por filtración de líquidos desde los vasos sanguíneos, típicas en la retinopatía diabética.
- **Hemorragias**: manchas rojas en la retina. Su forma puede variar; por ejemplo, en forma de llama si son más superficiales o redondeadas si están en capas más profundas.
- **Edema macular**: inflamación en la zona de la mácula. Puede sospecharse si desaparece el reflejo foveal o si hay exudados en forma circular.

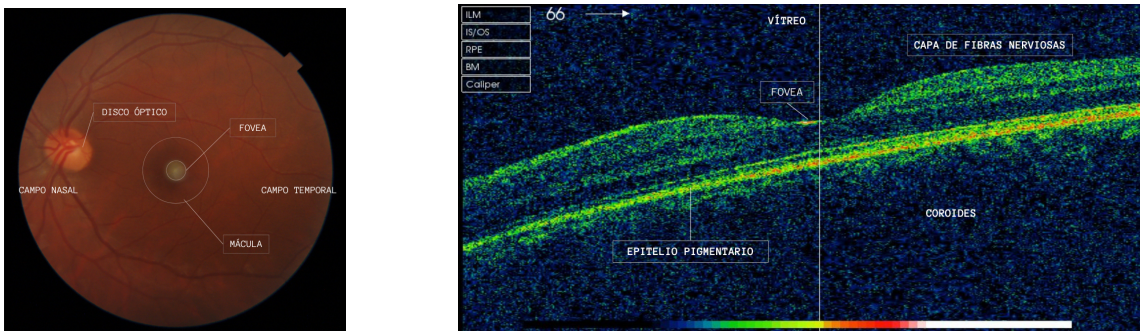
- **Excavación del nervio óptico**: cuando la zona central del nervio óptico está muy hundida, lo cual puede ser un signo de glaucoma [68].

Por otro lado, la imagen OCT (Tomografía de Coherencia Óptica) es como una ecografía pero con luz, que permite ver un corte en profundidad de la retina [69]. Muestra las distintas capas internas del tejido retiniano con gran detalle. Las partes más densas o reflectantes, como el epitelio pigmentario, se ven más brillantes, mientras que el líquido o las zonas vacías aparecen oscuras.

Con esta técnica se puede observar:

- **Capa de fibras nerviosas (RNFL)**: aparece como una banda brillante en la parte superior de la imagen; su grosor es importante para detectar el glaucoma [65].
- **Fóvea**: se ve como una pequeña hendidura central; su pérdida de forma puede indicar presencia de líquido o tracción.
- **Edema macular quístico**: zonas oscuras en el interior de la retina, que representan bolsas de líquido. Es frecuente en enfermedades como la diabetes.
- **Drusas**: se ven como pequeñas elevaciones bajo la retina, causadas por acumulación de desechos celulares, típicas en la DMAE [70].
- **Membranas epirretinianas**: líneas brillantes sobre la superficie de la retina, que pueden deformar las capas internas si tiran de ellas.

La Figura 3.5 muestra una comparación entre dos tipos de imágenes del ojo, en las que se han anotado las principales estructuras anatómicas relevantes para el análisis clínico. A la izquierda (a), se observa una imagen fundus donde se identifican claramente el disco óptico, la mácula, la fóvea y las regiones nasales y temporales de la retina. A la derecha (b), la imagen OCT muestra una sección transversal de la retina en la que se señalan varias capas importantes como la capa de fibras nerviosas, el epitelio pigmentario, la fóvea, el vítreo o la coroides. Esta segmentación detallada facilita el análisis estructural y la detección de alteraciones específicas.



(a) Imagen fundus (b) Imagen OCT

Figura 3.5: Ejemplos de imágenes fundus y OCT con anotaciones de las principales estructuras visuales [9]

Capítulo 4

Metodología

En este capítulo se explica la metodología seguida para implementar el módulo desarrollado para la detección automática de enfermedades visuales.

4.1. Materiales

Esta sección está dedicada a los materiales que se utilizan en este trabajo.

4.1.1. Conjunto de datos de imágenes fundus y OCT

El conjunto de datos que se utiliza en este trabajo es una versión del publicado por Taimur Hassan et al., titulado *A Composite Retinal Fundus and OCT Dataset along with Detailed Clinical Markings for Extracting Retinal Layers, Retinal Lesions and Screening Macular and Glaucomatous Disorders*, y disponible en la plataforma Mendeley Data [9]. Este conjunto fue desarrollado por investigadores de la National University of Sciences and Technology (NUST) y Khalifa University con el objetivo de facilitar la investigación en análisis automático de imágenes oftalmológicas mediante técnicas de inteligencia artificial.

El conjunto incluye imágenes, todas anonimizadas y validadas clínicamente, de 105 pacientes, cada uno con datos correspondientes al ojo derecho y al ojo izquierdo. Para cada ojo se dispone de dos modalidades de imagen: una imagen fundus (fotografía del fondo de ojo) y una imagen OCT (tomografía de coherencia óptica). Las imágenes se encuentran en formato PNG y JPEG, y se acompañan de anotaciones clínicas en archivos CSV. Estas etiquetas incluyen la clase diagnóstica de cada paciente, permitiendo su uso en tareas de clasificación supervisada.

Las cinco clases que componen el conjunto de datos son las siguientes:

- **Degeneración macular asociada a la edad (DMAE)**: 12 pacientes
- **Coriorretinopatía serosa central (CSR)**: 5 pacientes
- **Edema macular (ME)**: 12 pacientes
- **Glaucoma**: 26 pacientes

- **Healthy (sanos)**: 50 pacientes

Como puede observarse, existe un claro desbalance en la distribución de clases. Mientras que la clase *Healthy* representa casi la mitad del total de pacientes, otras como CSR están muy poco representadas. Este desequilibrio puede afectar negativamente al rendimiento de los modelos de aprendizaje profundo, generando sesgos hacia las clases mayoritarias. A lo largo del desarrollo del sistema se han tenido en cuenta estrategias de validación que compensen en la medida de lo posible esta limitación.

Las imágenes están organizadas inicialmente en dos carpetas principales: Macula y OD. En la carpeta Macula se incluyen las imágenes correspondientes a las patologías con afectación macular (DMAE, ME y CSR), mientras que en OD se agrupan los casos orientados al estudio del nervio óptico, como el glaucoma. La clase *Healthy* está presente en ambas carpetas, lo que permite usar estos casos como grupo control tanto en tareas maculares como en el análisis del nervio óptico.

En la Figura 4.1 se representa de forma esquemática la estructura del conjunto de datos, indicando la distribución de carpetas, clases y modalidades disponibles para cada paciente.

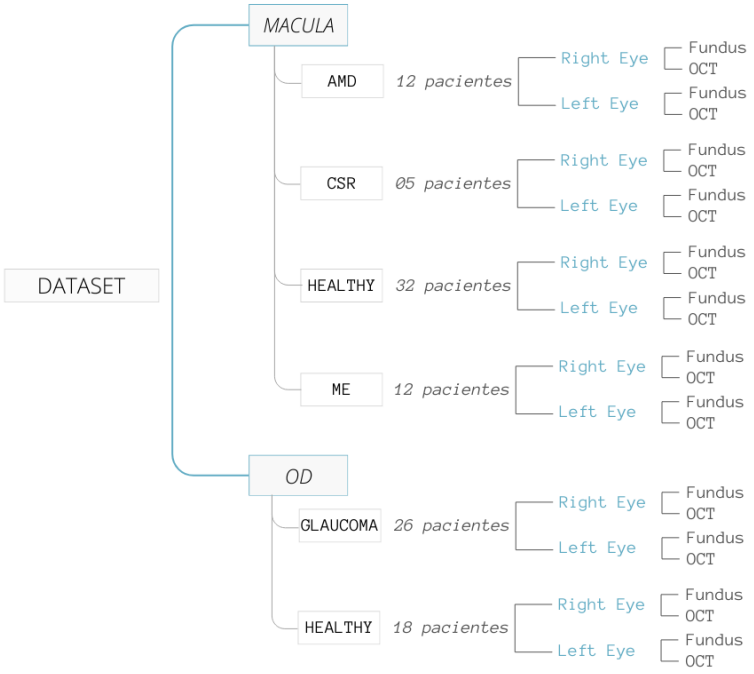


Figura 4.1: Estructura jerárquica del conjunto de datos. Las imágenes están divididas en dos carpetas principales (`Macula` y `OD`), cada una con pacientes clasificados por patología. Para cada paciente se incluyen imágenes fundus y OCT del ojo derecho e izquierdo.

4.2. Métodos

En esta sección, se detalla el procedimiento seguido a lo largo de la realización del trabajo. Este flujo de trabajo se sintetiza en el diagrama de la Figura 4.2. El proceso incluye el procesamiento y análisis de datos, seguido del modelado utilizando modelos preentrenados que se afinan para este proyecto específico utilizando el conjunto de imágenes de fondo de ojo y OCT.Posteriormente, se aplican estrategias de fusión de características con el objetivo de comprobar si el uso conjunto de ambas modalidades durante el entrenamiento permite obtener mejores resultados que cuando se emplean de forma individual y, finalmente, se realiza una evaluación exhaustiva de los resultados obtenidos. Cada uno de estos bloques será abordado en detalle en este trabajo.

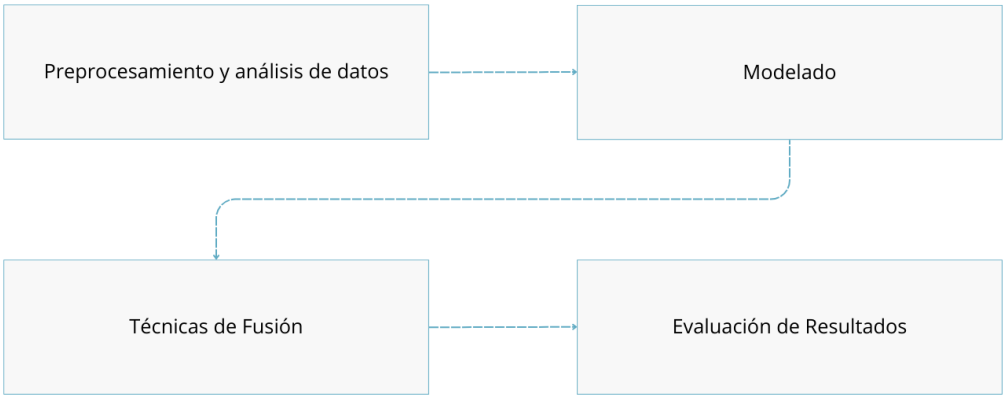


Figura 4.2: Flujo de trabajo o *pipeline* seguido en este proyecto.

4.2.1. Preprocesamiento y análisis de los datos

El proceso de preprocesamiento y análisis de datos es una fase fundamental que sienta las bases para una interpretación precisa de los resultados por parte de los modelos. En esta sección, se detalla cada etapa del proceso, incluidas las técnicas utilizadas para limpiar, transformar y explorar los datos, así como las herramientas empleadas para ello.

Como punto de partida, se realizó una exploración estructural del conjunto de datos para entender su organización interna. Para ello, se empleó la librería `os` de Python, que permite acceder a rutas de archivos, recorrer directorios y listar el contenido de las carpetas del sistema. A través de esta herramienta se inspeccionaron las carpetas principales `Macula` y `OD`, en las que se agrupan las imágenes por localización anatómica y clase diagnóstica.

Durante esta inspección inicial, se construyó un recuento del número de pacientes por clase, lo que permitió observar una distribución claramente desigual. Las frecuencias obtenidas fueron las siguientes:

- **Degeneración macular asociada a la edad (DMAE)**: 12 pacientes
- **Edema macular (ME)**: 12 pacientes

- **Coriorretinopatía serosa central (CSR)**: 5 pacientes
- **Glaucoma**: 26 pacientes
- **Healthy (sanos)**: 50 pacientes

Este primer análisis puso en evidencia un importante desbalance entre clases. Mientras que la clase *Healthy* representa prácticamente la mitad del conjunto de pacientes, otras como CSR están subrepresentadas. Este desequilibrio puede generar un sesgo durante el entrenamiento, favoreciendo a las clases con mayor presencia si no se aplican estrategias correctoras (como se comentó en la Sección 4.1.1).

Para cuantificar y visualizar de forma más precisa la distribución de clases, se construyó un `DataFrame` que recopilaba todas las etiquetas diagnósticas extraídas previamente. A partir de este, se contabilizó la frecuencia de aparición de cada clase y se normalizó dividiendo entre el total de muestras disponibles. Este proceso permitió obtener la proporción relativa de cada categoría diagnóstica dentro del conjunto, lo cual resultó clave para analizar el desbalance existente.

Con el objetivo de facilitar la interpretación de estos resultados, las proporciones calculadas se representaron mediante un gráfico circular, mostrado en la Figura 4.3. Esta representación visual permitió evidenciar, de forma clara, la sobrerrepresentación de la clase *Healthy* frente al resto.

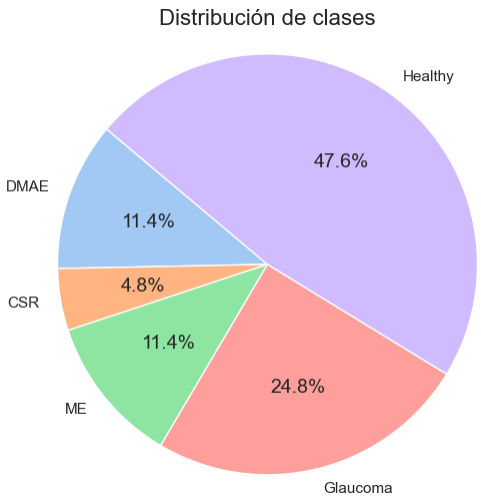


Figura 4.3: Distribución porcentual de pacientes por clase en el conjunto de datos.

Una vez analizada la distribución y verificada la coherencia de las etiquetas, se exportó el conjunto de datos final en formato `pickle`. Este formato, proporcionado por un módulo estándar de Python, permite serializar objetos complejos en forma binaria para su almacenamiento o transmisión, y posteriormente reconstruirlos en el mismo estado original. El archivo generado fue almacenado en *Google Drive*, lo que facilitó su acceso desde *Google Colab*, la plataforma en la

nube desarrollada por Google que permite ejecutar código Python en entornos de alto rendimiento, incluyendo el uso de GPU y TPU.

Con los datos organizados y accesibles, se procedió a construir una clase personalizada para la gestión estructurada del conjunto. Esta clase, denominada `CustomImageDataset`, fue implementada mediante la biblioteca `torch.utils.data.Dataset` de PyTorch, y su función principal era facilitar la carga dinámica de imágenes durante las fases de entrenamiento y validación del modelo.

Esta clase recorría la lista de pares `(ruta_imagen, etiqueta_str)` y se encargaba de abrir cada imagen utilizando la función `Image.open()` proporcionada por la biblioteca Pillow [71]. El resultado es un objeto de tipo `PIL.Image.Image`, que representa una imagen en memoria. Pillow es una librería de procesamiento de imágenes ampliamente utilizada en Python por su eficiencia y simplicidad, permitiendo realizar operaciones como conversión de formato, cambio de tamaño o ajustes de color de manera directa [71].

Además de abrir la imagen y convertirla a RGB, la clase se encargaba de transformar la etiqueta de texto (por ejemplo, "Glaucoma") a un valor numérico entero de manera incremental, utilizando un diccionario de mapeo. La correspondencia es la siguiente: 0: *DMAE*, 1: *CSR*, 2: *Healthy*, 3: *ME* y 4: *Glaucoma*. Este paso fue necesario para que las etiquetas fueran compatibles con la función de pérdida (*loss function*) del modelo, ya que muchas funciones de clasificación en PyTorch requieren que las clases estén codificadas como enteros consecutivos a partir de 0.

Este diseño modular y personalizado no solo permitió una integración directa con los `DataLoader` de PyTorch, sino que también facilitó la incorporación de transformaciones sobre las imágenes, como el redimensionado, la normalización o diversas técnicas de aumento de datos.

Por ello, dado que el conjunto de datos presentaba una cantidad limitada de muestras, se prestó especial atención a la prevención del sobreajuste. Para ello, se aplicaron estrategias de *data augmentation*, o aumento de datos, con el objetivo de incrementar artificialmente la diversidad del conjunto de entrenamiento. Esta técnica consiste en generar nuevas versiones de las imágenes originales mediante transformaciones controladas, permitiendo al modelo aprender de distintas variaciones de un mismo patrón visual sin necesidad de adquirir nuevas imágenes reales.

Inicialmente, las transformaciones se implementaron utilizando la biblioteca estándar de PyTorch. Las técnicas aplicadas fueron las siguientes:

- **Redimensionado (`Resize`)**: Todas las imágenes fueron reescaladas a una resolución fija (224x224) para asegurar que la entrada al modelo fuera uniforme, independientemente del tamaño original.
- **Conversión a tensor (`ToTensor`)**: Las imágenes, originalmente en formato PIL, fueron convertidas a tensores para poder ser utilizadas por los modelos en PyTorch.
- **Normalización (`Normalize`)**: Se aplicó una normalización de los valores

de píxel, utilizando la media y desviación estándar de la distribución de imágenes, con el fin de acelerar la convergencia del modelo.

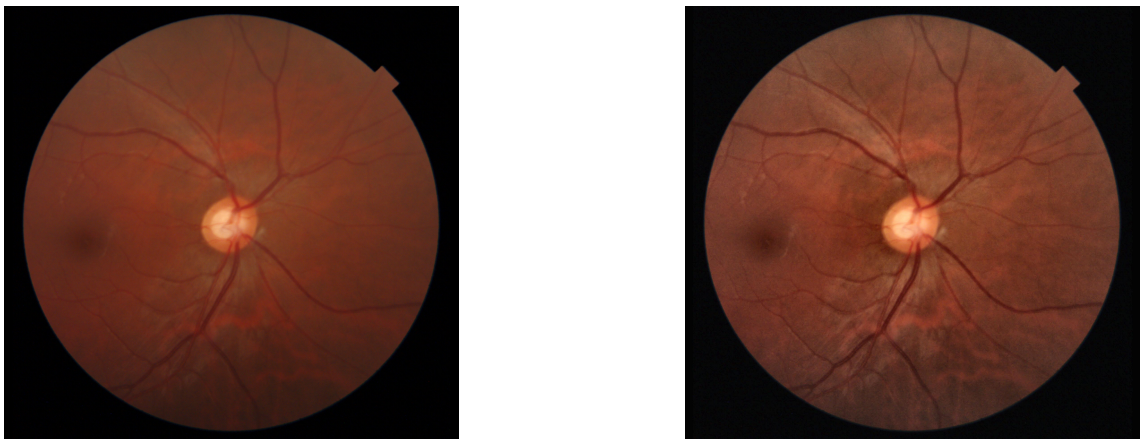
- **Rotación aleatoria**: Se introdujo una rotación aleatoria de los objetos dentro de un rango predefinido, permitiendo al modelo aprender a reconocer estructuras desde diferentes orientaciones.

- **Transformación `RandomAffine`**: Se aplicó una transformación afín con un ángulo de rotación máximo de 15 grados, lo que introduce pequeñas deformaciones geométricas en la imagen.

Sin embargo, durante la implementación de las estrategias de fusión multimodal, se detectó un problema importante: las transformaciones aleatorias aplicadas a cada imagen podían ser diferentes entre las dos modalidades, lo que resultaba en pares no alineados. Para resolverlo, se optó por utilizar la biblioteca `Albumentations` [72], especializada en tareas de visión por computador, que permite aplicar transformaciones idénticas de forma sincronizada sobre múltiples imágenes (por ejemplo, imagen y máscara o, en este caso, imagen fundus y su OCT correspondiente).

Entre las transformaciones más relevantes aplicadas con esta nueva biblioteca, destaca el uso del método **CLAHE (Contrast Limited Adaptive Histogram Equalization)** [73]. Esta técnica ajusta el contraste local de una imagen mediante la ecualización de histograma adaptativa en bloques, limitando la amplificación del ruido [73]. En imágenes fundus, su aplicación permite resaltar elementos clínicamente relevantes como drusas, exudados o microhemorragias, facilitando su interpretación tanto por humanos como por modelos automáticos.

La Figura 4.4 muestra un ejemplo comparativo entre una imagen fundus original y la misma imagen tras la aplicación de CLAHE. Se observa claramente una mejora en la visibilidad del detalle anatómico y patológico.



(a) Imagen fundus sin procesar (b) Imagen transformada

Figura 4.4: Comparacion entre imagen fundus sin procesar y tras la aplicación de CLAHE

Además de estas técnicas de transformación, también se experimentó con métodos adicionales de aumento de datos a nivel de conjunto. En concreto, se realizó un *oversampling manual*, replicando imágenes de clases minoritarias hasta triplicar el tamaño del conjunto de entrenamiento. Esta estrategia permitió balancear artificialmente las clases, igualando el número de muestras por categoría y reduciendo así el sesgo hacia las clases mayoritarias durante el aprendizaje.

Adicionalmente, se empleó la función `WeightedRandomSampler` de PyTorch, que permite ajustar la probabilidad de muestreo de cada clase en los lotes de entrenamiento. De esta forma, sin necesidad de duplicar físicamente las imágenes, se garantizó una mayor presencia de las clases menos representadas durante el entrenamiento, reforzando el equilibrio del conjunto.

Por último, para definir cómo se deben agrupar y preprocesar los datos antes de alimentarlos al modelo, se utilizó el módulo `DataLoader` de PyTorch, en combinación con la clase personalizada, explicada anteriormente, `CustomImageDataset`. Esta configuración permitió automatizar el proceso de carga por lotes (*batching*) y preparar de manera eficiente los tensores de entrada durante el entrenamiento.

En cada iteración, el `DataLoader` realiza las siguientes tareas:

1. Carga las imágenes correspondientes a un lote, accediendo a sus rutas previamente almacenadas.
2. Cada imagen es abierta y transformada en un objeto de tipo `PIL.Image.Image`, y posteriormente convertida a un tensor numérico, listo para ser procesado por el modelo.
3. Se aplican, de forma coherente, las transformaciones definidas previamente, como redimensionado, normalización o aumentos de datos, en función del modo (entrenamiento o validación).
4. Las etiquetas asociadas a cada imagen se transforman desde cadenas de texto a valores enteros, utilizando el mapeo definido en el preprocesamiento (por ejemplo, "Glaucoma"$\rightarrow$ 4).
5. Finalmente, el `DataLoader` devuelve un diccionario o lote que contiene dos elementos clave: los tensores de imagen y las etiquetas codificadas, que son usados directamente por el modelo durante la fase de entrenamiento o evaluación.

4.2.2. Modelado

La etapa de modelado abarca tanto la elección de los extractores de características como el diseño del clasificador encargado de realizar la predicción final. En este trabajo, se decidió realizar una comparación sistemática entre dos modelos preentrenados de tipo *Vision Transformer*: **DINOv2** y **CLIP**, utilizados como base para todos los experimentos.

Ambos modelos se utilizaron como extractores de características congelados (*frozen*) y fueron aplicados tanto a imágenes fundus como a imágenes OCT, con el

objetivo de evaluar su rendimiento en ambos contextos. Posteriormente, las representaciones extraídas se introdujeron en un clasificador propio de tipo `MLP` (Multilayer Perceptron) para obtener la predicción final.

Algunas diferencias clave entre estas dos arquitecturas se listan a continuación:

1. **Preentrenamiento:** DINOv2 ha sido entrenado de forma autosupervisada exclusivamente con imágenes, mientras que CLIP utiliza aprendizaje contrastivo sobre 400 millones de pares imagen-texto. Esto dota a CLIP de una capacidad semántica más amplia, mientras que DINOv2 se centra en estructuras visuales puras.
2. **Objetivo de entrenamiento:** CLIP aprende a alinear representaciones visuales con texto, lo que permite que sus embeddings capturen relaciones conceptuales más amplias. Por su parte, DINOv2 se centra en maximizar la invariancia de las representaciones sin supervisión directa.
3. **Robustez y transferencia:** DINOv2 genera representaciones altamente robustas sin necesidad de anotaciones, lo que lo hace ideal en contextos con escasez de datos etiquetados. CLIP, en cambio, es más sensible al contexto semántico y puede captar detalles de alto nivel, en particular en imágenes médicas.
4. **Eficiencia en tareas de visión médica:** En pruebas previas, DINOv2 ha demostrado buena generalización en imágenes de retina y OCT sin ajuste fino, mientras que CLIP ofrece un mayor rendimiento en tareas que requieren interpretación conceptual de patrones visuales complejos.

Adicionalmente, ambas arquitecturas se integraron en un mismo sistema de clasificación mediante un módulo común de tipo *MLP*, cuya arquitectura es la siguiente:

- Una primera capa `Linear` que proyecta las características extraídas por el modelo al espacio latente de dimensión `hidden` (768). Esta proyección inicial permite adaptar el espacio vectorial al tamaño requerido por las siguientes capas.
- Activación no lineal `GELU`, que introduce suavidad en la activación y conserva información de valores negativos pequeños. A diferencia de ReLU, GELU permite transiciones más suaves, lo que mejora el aprendizaje en tareas complejas como clasificación médica [74].
- Capa de `Dropout` con probabilidad 0.5. Esta técnica de regularización apaga aleatoriamente un porcentaje de las neuronas en cada paso de entrenamiento, reduciendo así el riesgo de sobreajuste [75].
- Capa `Linear` intermedia que reduce la dimensionalidad al 50 % del tamaño original (`hidden/2`), permitiendo una representación más compacta de las características.
- `LayerNorm`, una técnica de normalización que estabiliza y acelera el entrenamiento al mantener constante la distribución de activaciones dentro de cada muestra [76].

- Segunda activación `GELU` seguida de otra capa `Dropout` (0.5), que refuerzan la capacidad no lineal del modelo y su robustez frente al sobreajuste.
- Capa de salida `Linear` que proyecta al espacio de cinco dimensiones, correspondiente al número total de clases. Esta capa proporciona las puntuaciones (logits) que luego se utilizarán para calcular la función de pérdida.

Esta arquitectura permitió comparar ambos extractores bajo las mismas condiciones, facilitando un análisis justo y controlado de su rendimiento. En la Figura 4.5 se muestra el esquema de la arquitectura de la MLP

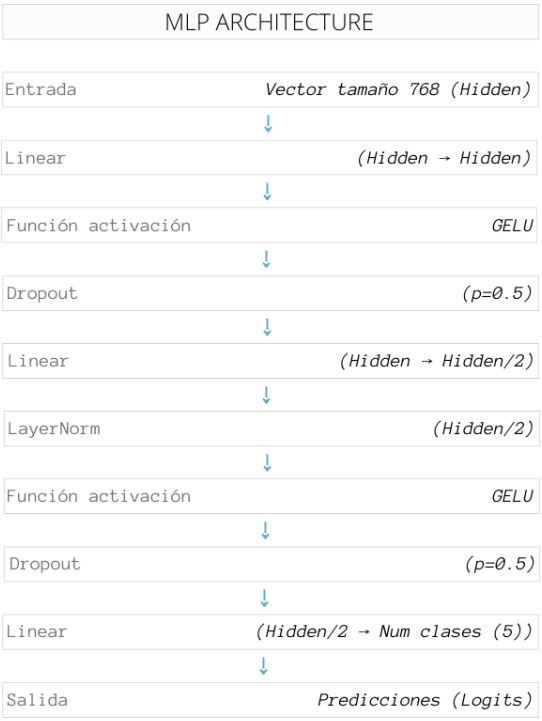


Figura 4.5: Arquitectura general de la red MLP

Una vez escogidos los modelos de extracción de características (DINOv2 y CLIP) y definido el clasificador MLP, se procedió a su entrenamiento. Para ello, fue necesario establecer previamente los hiperparámetros que configuran el comportamiento del entrenamiento. A continuación, se describen los hiperparámetros más relevantes utilizados:

- `batch_size`: Determina el número de muestras que se procesan simultáneamente en cada paso del entrenamiento.
- `num_epochs`: Indica el número total de épocas de entrenamiento. Cada época equivale a un paso completo por todos los datos del conjunto de entrenamiento.

Para la optimización del modelo se utilizó el optimizador *AdamW* [77], una variante del optimizador Adam que incorpora de forma explícita un término de de-

caimiento de pesos. Esta versión mejora el comportamiento de generalización y es especialmente útil en modelos de arquitectura Transformer, como los utilizados en este trabajo. La optimización estocástica subyacente permite ajustar los parámetros del modelo utilizando pequeños lotes de datos, en lugar de procesar el conjunto completo en cada iteración, lo que reduce los costes computacionales y acelera la convergencia [77].

Los hiperparámetros configurados para el optimizador AdamW fueron:

- `learning_rate`: Establece la tasa de aprendizaje inicial del optimizador *AdamW*.
- `weight_decay`: Este parámetro controla la cantidad de decaimiento de peso en el optimizador Adam que se aplica a todos los pesos de la red, excepto los pesos de sesgo y los pesos de la capa normalización [78].

El entrenamiento del sistema completo se llevó a cabo utilizando PyTorch, una biblioteca de aprendizaje profundo de código abierto ampliamente adoptada en la comunidad científica. Se implementaron ciclos personalizados de entrenamiento y validación, que permitieron mayor control sobre la lógica del modelo, el cálculo de métricas y la visualización de resultados.

4.2.3. Técnicas de fusión de características

Con el objetivo de aprovechar la información complementaria que aportan las imágenes *fundus* y *OCT*, en este trabajo se han evaluado diferentes estrategias para combinar ambas modalidades. En concreto, se han implementado y comparado tres enfoques principales: *early fusion*, *intermediate fusion* y *late fusion*. Cada una de estas estrategias integra la información en una etapa distinta del modelo, y su eficacia puede variar en función del extractor utilizado y del tipo de datos.

Para poder llevar a cabo estas estrategias de fusión, fue necesario modificar la clase del dataset desarrollada previamente, denominada *CustomImageDataset* (descrita en la sección 4.2.1), con el fin de adaptarla a una estructura que acepta entradas multimodales en forma de tripletas `(fundus_path, oct_path, label)`. De este modo, se garantiza que durante el entrenamiento y la inferencia se disponga de ambas imágenes para cada muestra del conjunto de datos.

En la Figura 4.6 se ilustra el esquema de la arquitectura utilizada para la estrategia de *early fusion*. Como se observa, las imágenes de ambas modalidades (OCT y fundus) se combinan al principio del proceso mediante concatenación a nivel de canales, generando una única entrada multimodal.

Esta imagen fusionada se introduce directamente en un extractor de características visuales, que puede ser DINOv2 o CLIP, compartido para ambas fuentes. A partir de ahí, se utiliza el token *[CLS]* como representación global de la imagen, que se pasa posteriormente por una red MLP para obtener las predicciones finales. Esta integración temprana permite al modelo aprender representaciones conjuntas desde las primeras capas, lo que puede facilitar la detección de patrones cruzados entre las modalidades.

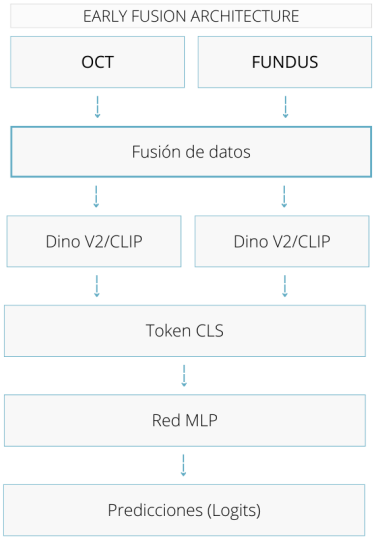


Figura 4.6: Arquitectura del modelo bajo la estrategia de *early fusion*.

Por otro lado, en la estrategia de *intermediate fusion*, la combinación entre ambas modalidades no se realiza directamente sobre las imágenes, sino a nivel de características. Como se observa en la Figura 4.7, cada imagen (fundus y OCT) se introduce por separado en su propio extractor de características (DINOv2 o CLIP). Esto permite que cada extractor aprenda representaciones específicas de su modalidad sin interferencias tempranas.

Posteriormente, las salidas obtenidas se proyectan a un espacio común mediante capas lineales (bloques de proyección) con el objetivo de igualar sus dimensiones y facilitar su integración. Una vez alineadas, ambas representaciones se concatenan y se introducen en un clasificador MLP, que se encarga de generar las predicciones finales.

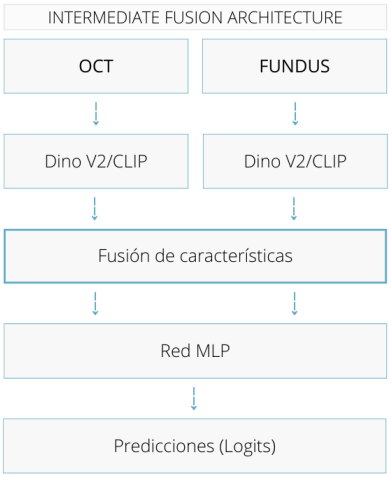


Figura 4.7: Arquitectura del modelo con fusión intermedia (*intermediate fusion*).

Por último, en la estrategia de *late fusion*, las modalidades fundus y OCT se procesan de forma completamente independiente a lo largo de todo el pipeline. Como se muestra en la Figura 4.8, cada imagen se introduce en su propio extractor de características (DINOv2 o CLIP), y posteriormente, las representaciones generadas se clasifican por separado mediante una red MLP individual.

Las predicciones obtenidas de cada ruta (fundus y OCT) se combinan únicamente en la etapa final del modelo. Para ello, se emplea una técnica sencilla como el promedio de los logits para obtener una única predicción. Esta aproximación permite mantener la independencia total entre ambas modalidades, aunque en la práctica puede perder parte de la riqueza que aporta una integración más temprana de la información.

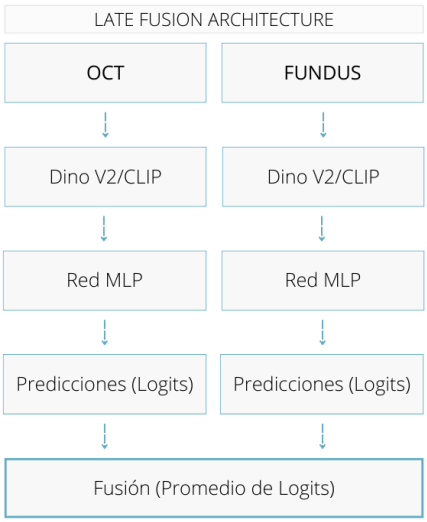


Figura 4.8: Arquitectura del modelo con estrategia de *late fusion*.

Cada una de estas configuraciones fue implementada utilizando ambos extractores (DINOv2 y CLIP), y sus resultados fueron comparados en base a métricas como *precisión, recall, F1-score y AUC*.

Capítulo 5

Experimentos y resultados

Este capítulo aborda el diseño experimental y los resultados obtenidos por el sistema de detección de anomalías en fotografías de fondo de ojo y OCT, indagando en los resultados individuales de cada modelo y de la fusión de datos, así como en una comparación de los resultados de cada uno.

5.1. Diseño experimental

El diseño experimental de este trabajo incluye todas las decisiones técnicas y configuraciones implementadas para desarrollar y evaluar los modelos de clasificación automática de enfermedades oculares mediante imágenes *fundus* y *OCT*. Este diseño se ha llevado a cabo utilizando una GPU T4 en el entorno de ejecución de Google Colab, lo que ha permitido realizar los experimentos con eficiencia computacional y tiempos de entrenamiento razonables.

La configuración general puede dividirse en dos bloques principales: el preprocesamiento de datos y la estrategia de entrenamiento de los modelos.

- **Preprocesamiento de datos**
  - Las imágenes de entrada se redimensionaron a una resolución de `224x224 píxeles`, valor recomendado para modelos preentrenados como DINOv2 y CLIP, lo que garantiza un formato uniforme de entrada.
  - Se aplicaron transformaciones básicas como `Resize`, `Normalize` y `ToTensor`. Además, debido a la limitada cantidad de datos, se evaluaron distintas estrategias de *data augmentation*, incluyendo rotaciones aleatorias, variaciones de brillo y contraste, transformaciones afines, y la técnica `CLAHE` para resaltar detalles como drusas en las imágenes fundus [73]. No obstante, y pese a los beneficios observados, se concluyó que la técnica más efectiva para mejorar el balance de clases y reducir el sobreajuste fue el uso del `WeightedRandomSampler`, que ajusta la frecuencia de muestreo durante el entrenamiento en función

de la clase, garantizando que el modelo reciba ejemplos equilibrados en cada lote (ver Sección 4.2.1).

- Para facilitar la separación de los datos entre entrenamiento y test, se implementó un script que divide el conjunto de pacientes en una proporción 80/20, asegurando que todas las imágenes (fundus y OCT) de un mismo paciente estén en el mismo subconjunto.

- **Modelado**

  - Se utilizaron dos extractores de características: DINOv2 y CLIP, aplicados de forma separada a imágenes *fundus* y *OCT*, generando así cuatro configuraciones principales. Todas las representaciones extraídas se pasaron por un mismo clasificador MLP, cuya arquitectura se detalla en la Sección 4.2.2.

  - Para garantizar una evaluación equilibrada, se mantuvo constante la configuración de entrenamiento para todos los modelos. A continuación, se listan los valores empleados:

    - `batch_size = 32`: Valor elegido para asegurar un equilibrio entre uso de memoria y velocidad de entrenamiento.

    - `num_epochs = 250`: Se seleccionó un número de 250 épocas completas como punto de partida común, permitiendo la comparación entre configuraciones.

  - El optimizador utilizado fue **AdamW**, aplicado de forma uniforme en todos los experimentos, con los siguientes hiperparámetros:

    - `weight_decay = 1e-5`: Penalización para reducir el sobreajuste aplicando regularización L2 a los pesos del modelo.

    - Se definieron dos tasas de aprendizaje independientes:

      - `lr_encoder = 1e-5`: Tasa de aprendizaje más baja para el extractor de características (DINOv2 o CLIP), dado que se emplean pesos preentrenados que no deben alterarse bruscamente.

      - `lr_decoder = 1e-3`: Tasa de aprendizaje más alta para el MLP, ya que esta parte del modelo se entrena desde cero y requiere mayor adaptación.

      Estas tasas de aprendizaje, junto con el valor de `weight_decay`, se han mantenido constantes para todas las configuraciones y experimentos realizados, con el objetivo de facilitar una comparación justa del rendimiento de cada modelo.

  - Como función de pérdida se empleó **CrossEntropyLoss**, la cual es adecuada para tareas de clasificación multiclase al comparar las probabilidades predichas con las etiquetas verdaderas codificadas como enteros [79].

  - Los modelos específicos utilizados en este trabajo fueron: `facebook/dinov2-base` para DINOv2 [80] y `openai/clip-vit-base-patch32` para CLIP [81]. Ambos han sido cargados a través de la librería `transformers` de Hugging Face.

- **Fusión de datos**

  - Se han evaluado tres estrategias de combinación de imágenes fundus y OCT: *early fusion*, *intermediate fusion* y *late fusion*, como se describe en la Sección 4.2.3.

  - En la estrategia de **early fusion**, ambas imágenes se concatenan canal a canal, generando una imagen de 6 canales que se introduce directamente en el encoder. Para ello, fue necesario modificar la configuración del modelo, ajustando el parámetro `config.in_channels` a 6 para aceptar esta entrada no estándar. Esta técnica permite capturar interacciones tempranas entre ambas modalidades, aunque puede resultar sensible a diferencias de escala o alineación.

  - En la **intermediate fusion**, las imágenes fundus y OCT se procesan de forma separada mediante sus respectivos extractores. Posteriormente, las representaciones latentes generadas por cada uno se proyectan a un espacio común y se concatenan antes de ser introducidas en el clasificador MLP. Esta estrategia busca equilibrar especialización y fusión temprana, conservando cierta independencia entre modalidades.

  - En la **late fusion**, cada modalidad se procesa de forma completamente independiente hasta obtener su predicción final. Las salidas se combinan posteriormente en la etapa de decisión, permitiendo máxima modularidad y tolerancia a errores en una de las fuentes.

- **Evaluación de los resultados**

  La evaluación de los resultados se ha realizado a partir de métricas estándar en problemas de clasificación multiclase, con el objetivo de analizar el rendimiento general y específico del modelo. A diferencia de los modelos de detección de objetos, en los que se utilizan métricas como *Intersection over Union (IoU) o Average precision (AP)* [82], este trabajo se enfoca en una tarea de clasificación de imágenes médicas, por lo que se han utilizado métricas basadas en etiquetas: *precisión, recall, F1-score y el área bajo la curva ROC (AUC)* (Ver Sección 2.6).

  Estas métricas permiten evaluar no solo la capacidad global del modelo para predecir correctamente las clases, sino también su comportamiento específico con cada una de las enfermedades. El cálculo de estas métricas se ha realizado tanto a nivel global como a nivel de clase, para identificar aquellas patologías que presentan mayor dificultad para el modelo.

  Para ello, se parte del conjunto de test y se recogen las predicciones del modelo para cada imagen. Posteriormente, se comparan con las etiquetas verdaderas para obtener los valores de verdaderos positivos, falsos positivos y falsos negativos por clase. Con estos datos, se calcula la precisión, el

recall, F1-score y el AUC (Ver sección 2.6).

Además del análisis global, se ha implementado una evaluación clase por clase para examinar en detalle el rendimiento del modelo en cada una de las cinco categorías del conjunto de datos: DMAE, CSR, Healthy, Edema macular y Glaucoma. Esta evaluación más granular resulta fundamental en el contexto médico, donde no todas las enfermedades tienen el mismo grado de dificultad de detección, ni están representadas de forma equilibrada.

Por otro lado, también se ha generado la matriz de confusión correspondiente para cada configuración de modelo, lo que permite visualizar de forma clara las confusiones más frecuentes entre clases. Este tipo de análisis ayuda a detectar casos concretos en los que el modelo tiende a fallar, como la confusión entre edema macular y CSR, o entre glaucoma y sujetos sanos.

5.2. Resultados

5.2.1. Resultados de entrenamiento

A continuación se presentan los resultados obtenidos durante el entrenamiento de las distintas configuraciones evaluadas, diferenciadas por modelo (DINOv2 y CLIP) y tipo de entrada (individual o fusión). En cada caso, se incluye la evolución de la función de pérdida y se comenta el comportamiento observado durante el proceso.

DINOv2 – Fundus individual

En esta configuración, el modelo fue entrenado exclusivamente con imágenes fundus empleando el extractor de características `facebook/dinov2-base`. Si bien el entrenamiento se había planificado para un total de 250 épocas, el proceso finalizó de forma anticipada en torno a la época 50 como consecuencia de la estrategia de *early stopping*, al no observarse mejoras significativas en la pérdida de validación tras varias iteraciones consecutivas.

Durante el entrenamiento, se implementó un mecanismo de guardado automático que almacenaba la versión del modelo con mejor rendimiento en validación, utilizando la función `torch.save()`. Esto garantizó que al finalizar el proceso se conservaran los pesos óptimos obtenidos a lo largo de las distintas épocas.

Tal como se representa en la Figura 5.1, la función de pérdida experimentó una reducción constante desde un valor inicial próximo a 1.60 hasta estabilizarse en torno a 0.25. Esta evolución progresiva, sin oscilaciones abruptas, evidencia un proceso de aprendizaje estable y eficiente. Además, el modelo mostró buena capacidad de generalización en esta configuración, lo que sugiere que las representaciones extraídas a partir de imágenes fundus son informativas y adecuadas para la tarea de clasificación propuesta.

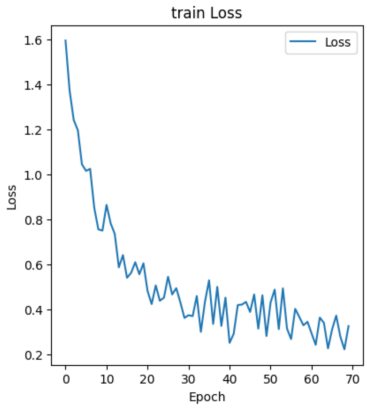


Figura 5.1: Pérdida durante el entrenamiento con fundus y DINOv2.

DINOv2 – OCT individual

En este experimento, se entrenó el modelo utilizando únicamente imágenes *OCT*, manteniendo el extractor de características `facebook/dinov2-base`. Como en el resto de configuraciones, el número de épocas máximo fue de 250, aunque el entrenamiento finalizó antes por la activación del mecanismo de *early stopping*, que detuvo el proceso en torno a la época 70 tras no observarse mejoras significativas en la función de pérdida.

Siguiendo el protocolo establecido para todos los experimentos, se utilizó la función `torch.save()` para almacenar automáticamente el modelo que mejor desempeño ofrecía en validación.

La Figura 5.2 muestra la evolución de la pérdida durante el entrenamiento. Se observa una disminución progresiva desde un valor inicial cercano a 1.85 hasta estabilizarse alrededor de 0.40. Aunque la convergencia se alcanzó de forma ligeramente más lenta que en el caso fundus, la tendencia descendente fue clara y sin irregularidades. Esta ligera diferencia puede deberse a la naturaleza más compleja de las imágenes OCT, que presentan estructuras internas más difíciles de interpretar que en las retinografías.

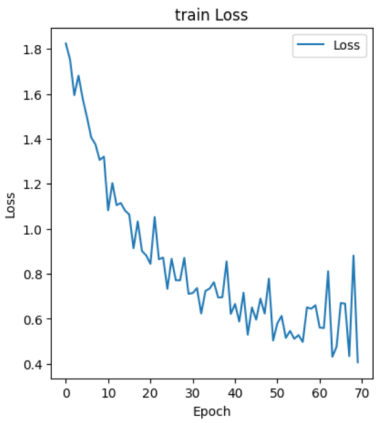


Figura 5.2: Pérdida durante el entrenamiento con OCT y DINOv2.

DINOv2 – Early fusion

En esta configuración, se empleó una estrategia de fusión temprana (*early fusion*), en la que ambas modalidades de imagen (fundus y OCT) fueron concatenadas a nivel de canal para formar una única entrada de seis canales. Para hacer esto posible, fue necesario modificar el valor del parámetro `config.in_channels = 6` en el encoder del modelo *DINOv2*, adaptando así su arquitectura para aceptar entradas multicanal. Esta técnica permite que el extractor de características aprenda desde las capas iniciales las correlaciones espaciales entre ambas modalidades, aprovechando patrones texturales y morfológicos complementarios.

El modelo fue entrenado durante las 250 épocas definidas, ya que no se activó el criterio de *early stopping*. Como se muestra en la Figura 5.3, la evolución de la función de pérdida fue descendente de forma sostenida a lo largo del entrenamiento, sin oscilaciones abruptas ni indicios de sobreajuste. La pérdida comenzó en torno a un valor de 1.8 y descendió progresivamente hasta estabilizarse alrededor de 0.2 en las últimas épocas.

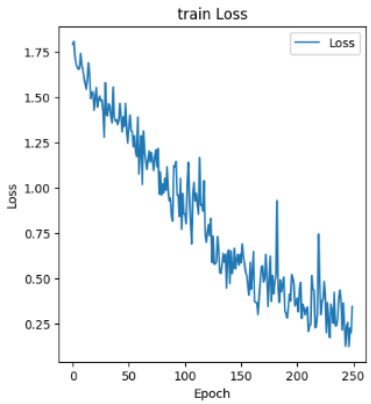


Figura 5.3: Pérdida durante el entrenamiento con fusión temprana y DINOv2.

DINOv2 – Intermediate fusion

En esta configuración, las imágenes fundus y OCT se procesaron por separado utilizando el mismo extractor de características, `DINOv2`. Una vez obtenidas las representaciones de ambas modalidades, se aplicaron capas lineales a cada una para proyectarlas a un espacio común. A continuación, estas representaciones se fusionaron y se utilizaron como entrada para la red neuronal MLP encargada de la clasificación.

Este enfoque permite que el modelo mantenga cierta especialización en el análisis de cada tipo de imagen (fundus y OCT), pero a la vez se beneficie de combinar la información de ambas modalidades antes de la toma de decisiones final. En este caso, el entrenamiento se completó en aproximadamente 90 épocas gracias al uso de *early stopping*, que detuvo el proceso al detectar que el modelo ya no mejoraba.

Tal como se muestra en la Figura 5.4, la pérdida de entrenamiento comenzó con

un valor alrededor de 1.75 y fue disminuyendo de forma constante hasta estabilizarse en torno a 0.15. Esta evolución progresiva y sin grandes oscilaciones indica que el modelo aprendió de forma estable y sin signos de sobreajuste. Al igual que en el resto de experimentos, se almacenó automáticamente el modelo con mejor rendimiento mediante `torch.save()`.

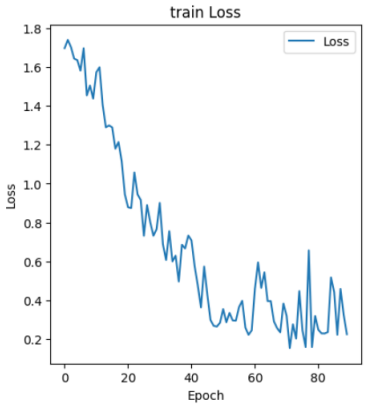


Figura 5.4: Pérdida durante el entrenamiento con fusión intermedia y DINOv2.

DINOv2 – Late fusion

En esta variante, se entrenaron dos modelos independientes, uno para imágenes fundus y otro para imágenes OCT, utilizando el extractor `DINOv2` en ambos casos. Cada uno de estos modelos produjo una predicción por separado, y ambas salidas se combinaron al final del pipeline para obtener la decisión final del sistema.

El entrenamiento se llevó a cabo durante las 250 épocas definidas por configuración, sin activar mecanismos de *early stopping*. Como se observa en la Figura 5.5, la función de pérdida mostró un descenso progresivo, comenzando en torno a 1.8 y alcanzando un valor por debajo de 0.4, aunque con un pequeño resalto al final. Esta evolución indica una convergencia estable, con una reducción continua del error a lo largo de las épocas, sin síntomas de sobreajuste.

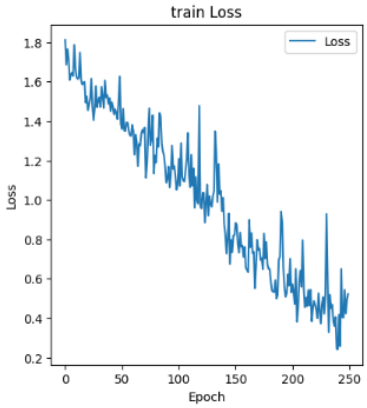


Figura 5.5: Pérdida durante el entrenamiento con fusión tardía usando DINOv2.

CLIP – Fundus individual

En este experimento, se usó el extractor preentrenado `openai/clip-vit-base-patch32` para entrenar el modelo con imágenes *fundus*. El entrenamiento mostró una convergencia más rápida en comparación con DINOv2, reflejando una mayor eficiencia del extractor en este tipo de imágenes.

Tal como se observa en la Figura 5.6, la función de pérdida experimenta una disminución pronunciada durante las primeras 25 épocas. A partir de ese punto, el descenso se vuelve más progresivo, y entre las épocas 35 y 50 se aprecian pequeñas oscilaciones sin mejoras significativas. Esta estabilización de la pérdida activa el criterio de parada anticipada, finalizando el entrenamiento en torno a la época 50.

La pérdida inicial se situaba en torno a 1.60, mientras que al final del entrenamiento descendió hasta aproximadamente 0.25, lo que indica una mejora significativa en el ajuste del modelo a los datos.

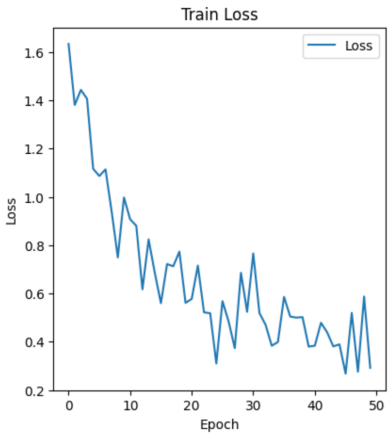


Figura 5.6: Pérdida durante el entrenamiento con fundus y CLIP.

CLIP – OCT individual

En esta configuración, se empleó únicamente la modalidad *OCT* como entrada para el modelo con el extractor `openai/clip-vit-base-patch32`. Como se aprecia en la Figura 5.7, la función de pérdida comenzó en torno a 1.8 y mostró un descenso progresivo hasta estabilizarse cerca de 0.4 en la época 70, momento en que se aplicó *early stopping*.

A lo largo del entrenamiento, el modelo presentó una evolución suave de la pérdida sin oscilaciones significativas, reflejando un aprendizaje estable. En comparación con el mismo experimento realizado con DINOv2, CLIP mostró una convergencia un poco más rápida y una pérdida final ligeramente inferior, lo cual sugiere una mejor adaptación de este extractor a las características estructurales de las imágenes OCT en este contexto.

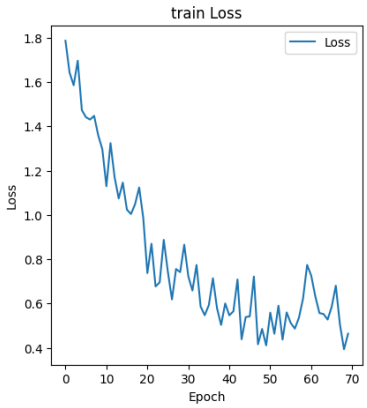


Figura 5.7: Pérdida durante el entrenamiento con OCT y CLIP.

CLIP – Early fusion

En esta configuración, las imágenes *fundus* y *OCT* se concatenaron a nivel de canal para formar una única entrada de seis canales. Al igual que en el caso de DINOv2, fue necesario ajustar el parámetro `config.in_channels = 6` para que el modelo pudiera aceptar este tipo de entrada combinada.

El entrenamiento mostró un descenso general de la función de pérdida a lo largo de las épocas, aunque con ligeras oscilaciones a partir del primer tercio del proceso. No obstante, el comportamiento general fue estable, sin indicios de sobreajuste. La función de pérdida partía de un valor aproximado de 2 y descendió hasta estabilizarse por debajo de 0.25.

El proceso concluyó tras 120 épocas debido al criterio de *early stopping*. Como en el resto de experimentos, tras cada iteración se almacenó automáticamente el modelo con mejor rendimiento en validación utilizando la función `torch.save()`.

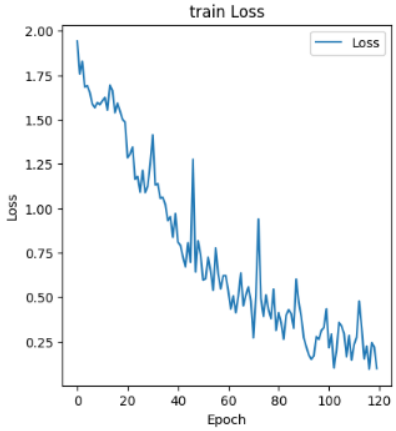


Figura 5.8: Pérdida durante el entrenamiento con fusión temprana usando CLIP.

CLIP – Intermediate fusion

En este experimento se utilizó la misma arquitectura empleada previamente con DINOv2, manteniendo el diseño de fusión intermedia, pero sustituyendo el extractor de características por el modelo CLIP. Las imágenes fundus y OCT fueron procesadas por separado a través del encoder, y sus representaciones se proyectaron a un espacio común mediante capas lineales. Posteriormente, ambas se fusionaron y se introdujeron en el MLP para realizar la clasificación.

Durante el entrenamiento se observó un descenso sostenido de la función de pérdida, con una evolución similar a la obtenida en el experimento equivalente con DINOv2. En concreto, la pérdida se redujo desde un valor inicial aproximado de 2 hasta en torno a 0.15, aunque al final hubiera un aumento de la pérdida provocando que el entrenamiento se detuviera de forma anticipada en torno a la época 50 mediante el mecanismo de *early stopping*, al no detectarse mejoras adicionales en la validación, tal y como ha ocurrido en otros experimentos.

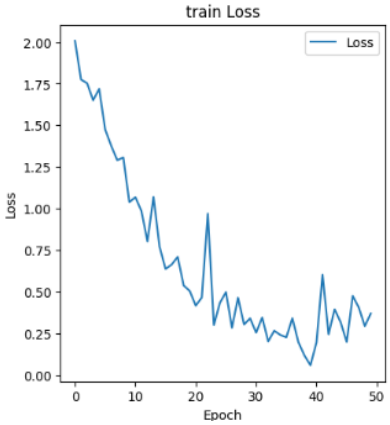


Figura 5.9: Pérdida durante el entrenamiento con fusión intermedia usando CLIP.

CLIP – Late fusion

En esta configuración, se empleó la misma arquitectura utilizada en la modalidad *late fusion* con DINOv2, sustituyendo únicamente el extractor de características por el modelo CLIP. Ambas modalidades (*fundus* y *OCT*) fueron procesadas por separado, y sus predicciones individuales se combinaron posteriormente para obtener la clasificación final.

Tal como se observa en la Figura 5.10, la función de pérdida experimentó una disminución progresiva a lo largo de las aproximadamente 55 épocas de entrenamiento, comenzando en un valor aproximado de 1.85 y finalizando en torno a 0.2. A pesar de la ligera variabilidad observada entre épocas, el comportamiento general de la curva fue descendente, lo cual refleja una capacidad de aprendizaje adecuada.

En comparación con la modalidad *late fusion* entrenada con DINOv2, el entrenamiento con CLIP presentó una menor estabilidad y oscilaciones más pronun-

ciadas. Esta diferencia sugiere que CLIP es más sensible a las perturbaciones durante el aprendizaje.

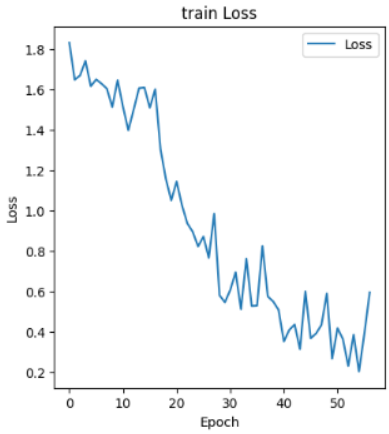


Figura 5.10: Pérdida durante el entrenamiento con fusión tardía usando CLIP.

5.2.2. Evaluación de resultados

La evaluación final de este trabajo tiene como objetivo valorar la eficacia de los modelos entrenados en el conjunto de test, aislado durante todo el proceso de entrenamiento, para clasificar enfermedades oculares a partir de imágenes *fundus* y *OCT*. Como línea base de este TFG, se plantea que la combinación de ambas modalidades, mediante técnicas de fusión multimodal, puede mejorar los resultados en comparación con el uso individual de cada fuente de imagen.

Para contrastar esta línea base, como ya se ha comentado anteriormente, se han llevado a cabo diferentes experimentos utilizando los modelos DINOv2 y CLIP, aplicados tanto de forma individual sobre cada modalidad como en configuraciones de fusión (*early*, *intermediate* y *late fusion*).

Para facilitar la comparación entre las distintas arquitecturas y configuraciones evaluadas, se han elaborado tres tablas que recogen las métricas de precisión obtenidas por cada modelo. Estas tablas permiten analizar el impacto del uso individual y combinado de las técnicas de fusión. Además, se incluyen matrices de confusión y ejemplos representativos que proporcionan una visión más detallada del comportamiento de los modelos en distintos escenarios de clasificación.

Comparación global por configuración

La Tabla 5.1 recoge los resultados macro para cada decoder y modelo, evaluados mediante métricas estándar: *precisión, recall, F1-score y AUC*.

Decoder	Modelo	Precisión	Recall	F1-Score	AUC
DINOv2	Fundus individual	**0.84**	**0.79**	**0.79**	**0.94**
	OCT individual	0.54	0.67	0.48	0.74
	Early fusion	0.70	0.72	0.69	0.90
	Intermediate fusion	0.78	0.68	0.70	0.90
	Late fusion	0.50	0.51	0.42	0.67
CLIP	Fundus individual	**0.80**	**0.85**	**0.81**	**0.95**
	OCT individual	0.79	0.82	0.74	0.92
	Early fusion	0.64	0.58	0.59	0.82
	Intermediate fusion	0.40	0.52	0.42	0.75
	Late fusion	0.54	0.46	0.48	0.86

Cuadro 5.1: Resultados ponderados por configuración y modelo en distintas métricas de evaluación. Se resaltan en negrita los mejores valores por métrica para cada modelo.

Como puede apreciarse en la Tabla 5.1, al analizar el rendimiento medio de cada modelo a lo largo de las distintas configuraciones, se observa que DINOv2 ofrece resultados más estables y competitivos. Aunque CLIP alcanza valores máximos muy elevados en determinadas configuraciones, DINOv2 muestra un comportamiento más robusto, especialmente en aquellas estrategias que implican la fusión de modalidades. Esto sugiere que el modelo se adapta mejor al uso conjunto de imágenes fundus y OCT, una capacidad especialmente relevante en tareas de clasificación multimodal.

Dentro de DINOv2, la configuración *fundus individual* es la que obtiene los mejores resultados globales, superando al resto de configuraciones en todas las métricas. A esta le siguen las estrategias de *early fusion* e *intermediate fusion*, que ofrecen un rendimiento equilibrado y coherente, confirmando que la combinación temprana o intermedia de características puede ser beneficiosa siempre que el modelo lo permita.

En el caso de CLIP, destaca también la modalidad *fundus individual*, con un F1-score de 0.81 y un AUC de 0.95. Estos resultados indican que el modelo se encuentra especialmente adaptado a las características superficiales que aportan las imágenes fundus, obteniendo un rendimiento muy competitivo sin necesidad de combinar múltiples fuentes de información.

Por el contrario, las configuraciones menos efectivas para ambos modelos corresponden a la estrategia de *late fusion*, donde la combinación de predicciones finales resulta menos eficaz. Esta limitación es especialmente notable en DINOv2, que alcanza un F1-score de apenas 0.42, y también se refleja en CLIP, con un valor de 0.48. Estos datos refuerzan la idea de que fusionar las modalidades en fases avanzadas puede limitar la capacidad del modelo para aprender representaciones conjuntas útiles.

Por otro lado, dado el desequilibrio existente entre las clases del conjunto de datos, resulta especialmente relevante analizar los resultados a partir de métricas ponderadas. En la siguiente tabla se presentan dichas métricas ponderadas.

Decoder	Modelo	Precisión	Recall	F1-Score	AUC
DINOv2	Fundus individual	**0.80**	**0.77**	**0.77**	**0.92**
	OCT individual	0.76	0.54	0.49	0.74
	Early fusion	0.65	0.65	0.64	0.84
	Intermediate fusion	0.72	0.69	0.69	0.85
	Late fusion	0.46	0.38	0.40	0.60
CLIP	Fundus individual	**0.88**	**0.85**	**0.81**	0.92
	OCT individual	0.86	0.81	0.80	**0.93**
	Early fusion	0.60	0.58	0.54	0.79
	Intermediate fusion	0.57	0.65	0.59	0.75
	Late fusion	0.60	0.62	0.59	0.76

Cuadro 5.2: Resultados ponderados por configuración y modelo en distintas métricas de evaluación. Se resaltan en negrita los mejores valores por métrica para cada modelo.

Como se observa en la Tabla 5.2, el rendimiento ponderado pone de manifiesto ciertas diferencias respecto a la evaluación macro. En particular, se evidencia una mejora en las configuraciones con menor representación de clases, como es el caso de las modalidades individuales de OCT, que ahora obtienen puntuaciones más competitivas, especialmente en el modelo CLIP.

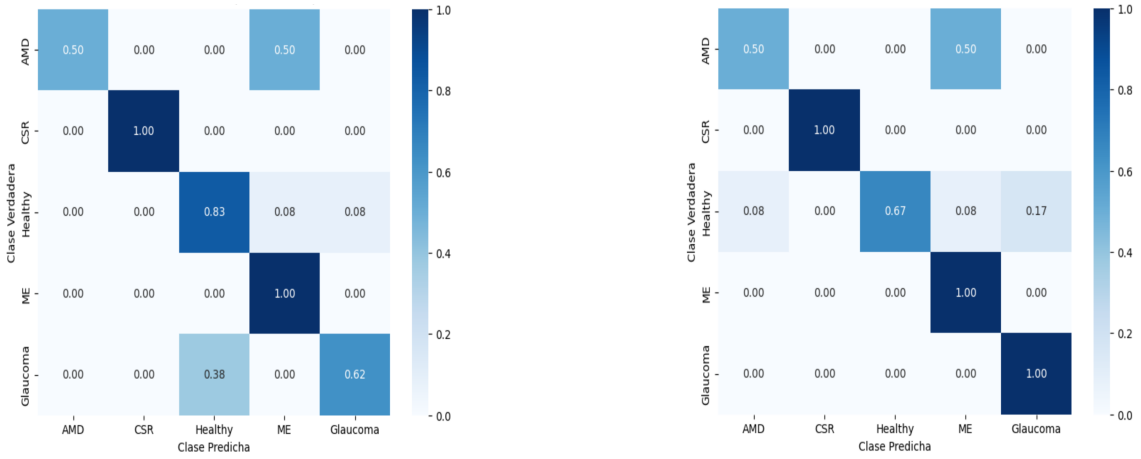
En términos generales, CLIP sigue destacando por su rendimiento en configuraciones individuales. En concreto, la modalidad *fundus individual* alcanza el F1-score más alto (0.81), junto con un AUC de 0.92, mientras que *OCT individual* ofrece un desempeño muy parecido, con un F1-score de 0.80 y el mayor AUC registrado (0.93). Esto sugiere que CLIP es capaz de extraer representaciones robustas tanto de imágenes fundus como de OCT por separado, lo que resulta especialmente relevante en contextos donde solo una modalidad está disponible.

Por su parte, DINOv2 muestra un comportamiento más equilibrado en configuraciones de fusión, especialmente en la estrategia de *intermediate fusion*, que obtiene valores cercanos a los máximos del modelo (F1-score de 0.69 y AUC de 0.85). Aunque la modalidad *fundus individual* continúa siendo la más destacada, la diferencia con las configuraciones multimodales se reduce en comparación con la evaluación macro, lo que indica una mayor sensibilidad del modelo a la distribución real de las clases.

Al igual que en el análisis anterior, la estrategia de *late fusion* vuelve a posicionarse como la menos efectiva, con F1-scores de 0.40 para DINOv2 y 0.59 para CLIP.

Matrices de confusión

A continuación, se presentan las matrices de confusión correspondientes a los mejores resultados obtenidos por cada modelo en configuración individual, así como la mejor configuración con fusión. Estas matrices permiten observar cómo se comportan los modelos en cada clase y dónde se concentran los errores de clasificación.



(a) DinoV2 (b) CLIP

Figura 5.11: Matrices de confusión de los modelos DinoV2 y CLIP con modalidad fundus individual

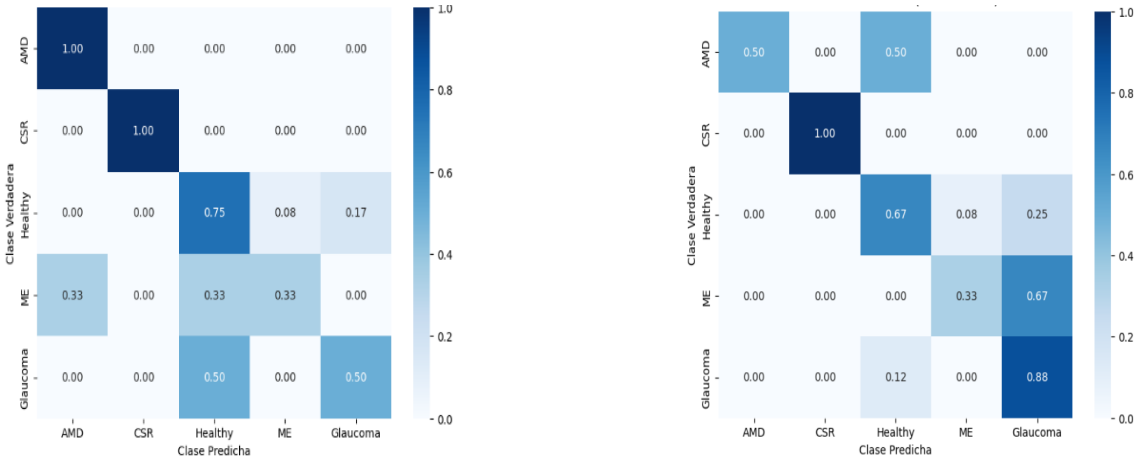
En la Figura 5.11a, el modelo DINOv2 logra una clasificación perfecta en las clases CSR y Edema Macular (ME), con una precisión del 100 %. También destaca su buen rendimiento en la clase Healthy (83 % de acierto), aunque presenta cierta confusión con ME y Glaucoma. El mayor problema aparece en la clase DMAE, donde solo el 50 % de las instancias se clasifican correctamente, siendo el resto etiquetado como ME, posiblemente debido a similitudes en los patrones visuales.

Por su parte, el modelo CLIP (Figura 5.11b) también muestra una clasificación perfecta en CSR, ME y Glaucoma. Sin embargo, el rendimiento en la clase Healthy es ligeramente inferior al de DINOv2, con un 67 % de aciertos, y una mayor dispersión de errores hacia DMAE y Glaucoma. En el caso de DMAE, se repite la tendencia observada con DINOv2, donde el modelo también divide sus predicciones entre DMAE y ME a partes iguales (50 % cada una), indicando una dificultad compartida.

Aunque ambos modelos logran buenos resultados generales en esta configuración, la comparación directa entre las dos matrices revela diferencias importantes. DINOv2 muestra una mejor capacidad para reconocer la clase Healthy, lo que puede ser relevante dada su representación frecuente en el conjunto de datos. Por otro lado, CLIP sobresale ligeramente en la identificación de Glaucoma y mantiene una clasificación perfecta en las mismas clases que DINOv2. Estas diferencias reflejan que, aunque ambos modelos son eficaces en tareas

individuales con imágenes fundus, cada uno tiende a especializarse en distintos patrones patológicos. La elección entre uno u otro puede depender del contexto clínico específico y del tipo de patologías prioritarias a detectar.

Además de las configuraciones individuales, algunas estrategias de fusión también lograron resultados competitivos, especialmente en el caso de DINOv2, donde tanto la *early fusion* como la *intermediate fusion* ofrecieron un rendimiento cercano al de los mejores modelos. A continuación, se muestran las matrices de confusión correspondientes a estas configuraciones, lo que permite profundizar en su comportamiento frente a las distintas clases.



(a) Early fusion

(b) Intermediate fusion

Figura 5.12: Matrices de confusión de los mejores modelos de fusión con DinoV2

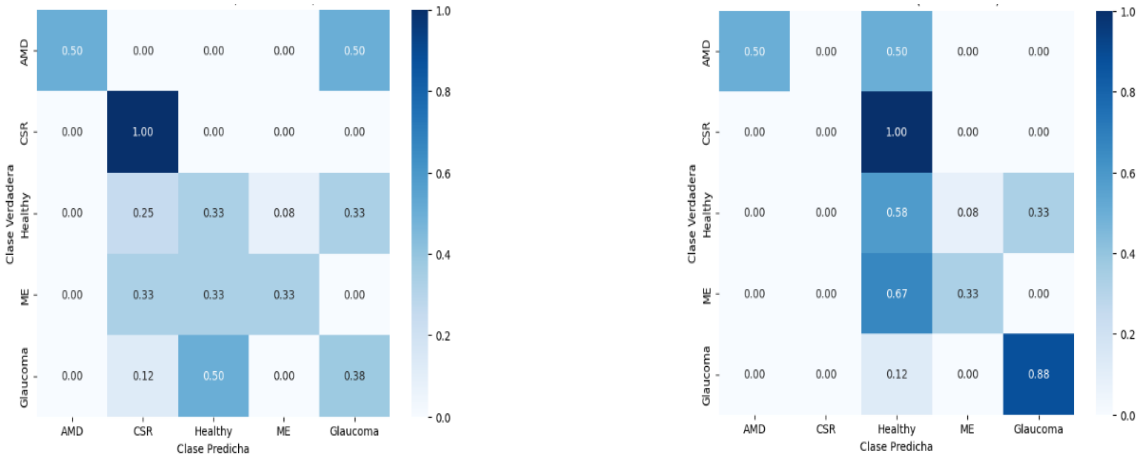
La Figura 5.12a muestra la matriz de confusión del modelo DINOv2 con *early fusion*. En este caso, se observan muy buenos resultados en las clases AMD y CSR, que son clasificadas correctamente en su totalidad. La clase Healthy alcanza un valor de acierto del 75 %, aunque presenta confusiones con las clases ME y Glaucoma. En cuanto a Glaucoma, el modelo la confunde en el 50 % de los casos con Healthy, lo que indica una cierta dificultad para distinguirla de una retina sana en algunos ejemplos.

Por su parte, la Figura 5.12b correspondiente a *intermediate fusion* presenta un patrón de clasificación algo diferente. Aunque el rendimiento en CSR y Glaucoma sigue siendo muy sólido, alcanzando valores cercanos al 100 %, el modelo muestra más errores en Healthy, que es confundida con Glaucoma en un 25 % de los casos. También se aprecia una ligera confusión en la clase AMD, que se divide equitativamente entre predicciones correctas y errores hacia Healthy.

Al comparar ambas configuraciones de fusión, puede concluirse que *early fusion* consigue una mayor homogeneidad en el reparto de aciertos entre clases, observando cómo clasifica mejor las clases minoritarias.

En el extremo opuesto, las configuraciones asociadas a la estrategia de *late fusion* han mostrado los peores resultados en ambos modelos. A continuación se

presentan las matrices de confusión correspondientes para DINOv2 y CLIP.



(a) DinoV2 (b) CLIP

Figura 5.13: Matrices de confusión de las configuraciones *late fusion* para DinoV2 y CLIP

En la Figura 5.13a, correspondiente al modelo DINOv2, se observa un rendimiento inconsistente, especialmente en clases como Healthy y Glaucoma, que presentan una alta tasa de error. Aunque CSR se clasifica correctamente en todos los casos, otras clases como AMD y Glaucoma muestran una división equitativa entre predicciones correctas e incorrectas. La clase Healthy es especialmente problemática, siendo confundida con CSR, ME y Glaucoma en un total del 66 % de los casos.

En el caso de CLIP (Figura 5.13b), la situación no mejora. Aunque la clase Glaucoma tiene una precisión prácticamente perfecta (88 %), el resto muestra errores significativos. Healthy y ME se clasifican con porcentajes de acierto muy bajos (58 % y 33 % respectivamente), y la clase CSR sufre una pérdida considerable de precisión, cayendo al 0 %, con confusión principalmente hacia Healthy. AMD repite el patrón de acierto parcial observado en DINOv2.

Al comparar ambas figuras, se aprecia que, si bien los dos modelos fallan en aspectos similares, CLIP parece manejar algo mejor la clase Glaucoma, mientras que DINOv2 muestra un reparto más homogéneo de errores. En cualquier caso, estas configuraciones confirman la debilidad de la estrategia de *late fusion*, que se basa en el promedio de las predicciones finales (logits) de los modelos individuales. Este enfoque, al no tener en cuenta la interacción directa entre las modalidades fundus y OCT, podría estar perdiendo información relevante para la clasificación, especialmente en clases más complejas.

Este análisis refuerza la idea de que fusionar modalidades en etapas anteriores del modelo, como ocurre en *early* e *intermediate fusion*, puede facilitar una mejor integración de la información y, por tanto, mejorar el rendimiento.

Análisis por clase: F1-score y AUC

Dado que se trata de un problema de clasificación multiclase con clases desbalanceadas, resulta especialmente relevante analizar el comportamiento del modelo en cada categoría. Para ello, se han empleado las métricas F1-score y AUC, que permiten evaluar conjuntamente la sensibilidad (*recall*) y la precisión de cada clase, siendo especialmente útiles en escenarios donde un mal desempeño en una sola categoría puede tener consecuencias clínicas relevantes.

Decoder	Modelo	DMAE	CSR	Healthy	EM	Glaucoma
DINOv2	Fundus individual	0.67	**1.00**	**0.80**	**0.75**	**0.78**
	OCT individual	0.55	0.84	0.67	0.81	0.85
	Early fusion	**0.80**	**1.00**	0.69	0.40	0.57
	Intermediate fusion	0.67	**1.00**	0.74	0.40	0.70
	Late fusion	0.67	0.29	0.38	0.40	0.38
CLIP	Fundus individual	0.50	**1.00**	**0.86**	**0.75**	0.94
	OCT individual	**0.75**	0.67	0.80	0.50	**1.00**
	Early fusion	0.50	**1.00**	0.67	0.40	0.36
	Intermediate fusion	0.00	0.00	0.70	0.57	0.84
	Late fusion	0.67	0.00	0.58	0.40	0.74

Cuadro 5.3: F1-score por clase para cada configuración y modelo. Se resaltan en negrita los mejores resultados por clase para cada modelo.

Tal y como se observa en el Cuadro 5.3, los mejores valores de F1-score por clase se distribuyen principalmente entre las configuraciones individuales, especialmente en el modelo DINOv2. En concreto, la configuración *Fundus individual* con DINOv2 destaca por obtener los mayores valores de F1-score en cuatro de las cinco clases: CSR (1.00), Healthy (0.80), EM (0.75) y Glaucoma (0.78), lo que evidencia un desempeño muy equilibrado para esta modalidad.

En el caso de CLIP, la estrategia *Fundus individual* también logra resultados sólidos, con valores destacados en CSR (1.00), Healthy (0.86) y EM (0.75). Por otro lado, la configuración *OCT individual* en CLIP obtiene el mejor valor para la clase DMAE (0.75) y mantiene una alta puntuación en Glaucoma (1.00), lo que indica una buena capacidad para detectar patologías estructurales.

Respecto a las estrategias de fusión, *Early fusion* en DINOv2 sobresale al alcanzar el mejor valor para DMAE (0.80) y mantener una puntuación perfecta en CSR (1.00), mostrando su efectividad para combinar información multimodal en fases tempranas. También se observa un rendimiento razonable en la clase Healthy (0.69), aunque cae notablemente en otras como EM y Glaucoma. En contraste, las estrategias de *Intermediate* y *Late fusion*, especialmente cuando se aplican a CLIP, muestran un descenso generalizado en la mayoría de las clases.

En el caso de la *DMAE*, se observa que la *early fusion* con DINOv2 alcanza el ma-

yor valor de F1-score (0.80), superando tanto a las configuraciones individuales como al resto de estrategias de fusión. Este resultado es coherente con la naturaleza de esta patología, ya que la combinación de las dos modalidades aporta información complementaria: mientras las imágenes fundus permiten detectar señales superficiales como drusas, las imágenes OCT ofrecen una visión transversal detallada que revela alteraciones estructurales, como el engrosamiento del epitelio pigmentario. Al combinar ambas representaciones desde el inicio del pipeline, el modelo es capaz de integrar estas características, mejorando así la capacidad de detección.

Decoder	Modelo	DMAE	CSR	Healthy	EM	Glaucoma
DINOv2	Fundus individual	0.91	**1.00**	**0.89**	**0.95**	**0.94**
	OCT individual	0.54	0.84	0.67	0.81	0.85
	Early fusion	**0.98**	**1.00**	0.79	0.87	0.85
	Intermediate fusion	0.96	**1.00**	0.80	0.87	0.86
	Late fusion	0.50	0.80	0.85	0.49	0.57
CLIP	Fundus individual	0.95	**1.00**	0.86	**0.96**	0.97
	OCT individual	0.94	0.95	**0.92**	0.72	**1.00**
	Early fusion	0.83	**1.00**	0.77	0.65	0.83
	Intermediate fusion	0.48	0.76	0.76	0.81	0.96
	Late fusion	**0.96**	**1.00**	0.63	0.86	0.83

Cuadro 5.4: AUC por clase para cada configuración y modelo. Se resaltan en negrita los mejores resultados por clase para cada modelo.

En cuanto al AUC (Cuadro 5.4), los mejores resultados se obtienen, en general, con configuraciones individuales. En DINOv2, la modalidad *Fundus individual* ofrece los valores más altos en la mayoría de clases, destacando especialmente en CSR (1.00), EM (0.95) y Glaucoma (0.94). CLIP también muestra un rendimiento excelente con *Fundus individual*, alcanzando su máximo en Glaucoma (0.97), EM (0.96) y CSR (1.00).

En cuanto al uso de técnicas de fusión, *Early* e *Intermediate fusion* con DINOv2 logran resultados competitivos en casi todas las clases, manteniendo valores altos en CSR, EM y Glaucoma. No ocurre lo mismo con CLIP, donde el rendimiento baja en algunas clases al aplicar estas estrategias, especialmente en *Intermediate fusion* para DMAE (0.48) o *Late fusion* en Healthy (0.63).

En general, el análisis sugiere que DINOv2 maneja mejor la combinación de modalidades, ya que logra mantener valores altos de AUC incluso cuando se integran ambas fuentes de información. Además, se confirma que las clases más difíciles de distinguir siguen siendo EM y, en menor medida, DMAE, donde los modelos tienden a obtener los resultados más bajos. Esto podría estar relacionado con una mayor variabilidad visual o una menor representación de estas clases en el conjunto de datos.

5.2.3. Análisis cualitativo

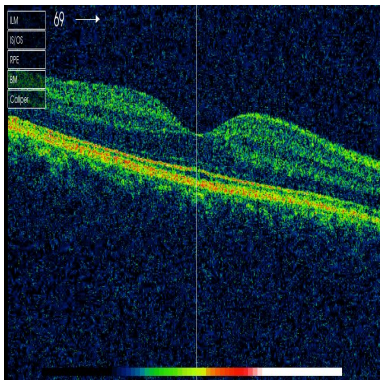
Para complementar el análisis cuantitativo, se han recopilado ejemplos visuales representativos que permiten observar cómo se comportan los modelos ante distintas imágenes. Este análisis cualitativo ayuda a comprender mejor tanto los aciertos como las limitaciones de las configuraciones evaluadas.

De acuerdo con las Tablas 5.3 y 5.4, las clases *Healthy* y *Glaucoma* han sido las más consistentemente clasificadas correctamente por la mayoría de los modelos, tanto en términos de F1-score como de AUC. Esto indica que las características visuales de estas categorías son más fácilmente reconocibles por los extractores y clasificadores utilizados, probablemente debido a patrones estructurales más definidos y a una mayor representación en el conjunto de datos.

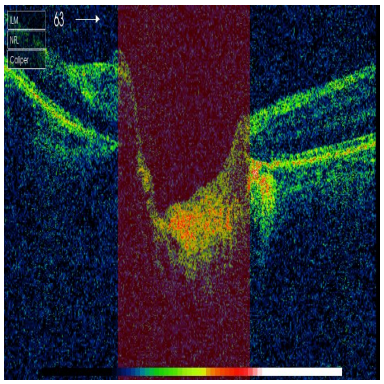
Aunque la clase *Edema Macular (EM)* también ha sido clasificada por todos los modelos en las pruebas, los resultados de precisión han sido más variables. Lo que sugiere que, aunque los modelos son capaces de identificar imágenes de EM hasta cierto punto, tienen más dificultades para evitar errores en esta categoría, probablemente por su similitud visual con otras patologías como DMAE o por una menor calidad y diversidad en las imágenes disponibles de esta clase.

(a) Fundus Healthy

(b) Fundus Glaucoma



(c) OCT Healthy

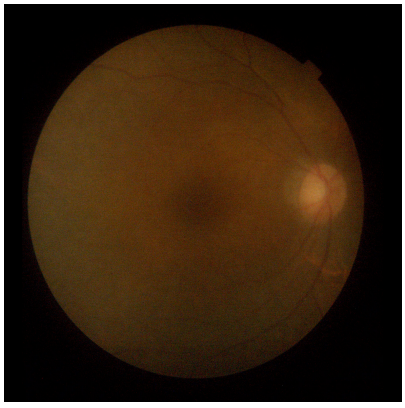


(d) OCT Glaucoma

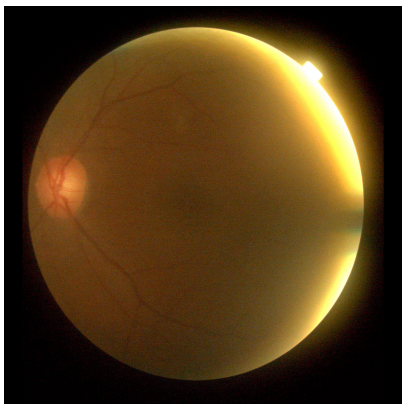
Figura 5.14: Ejemplos de imágenes correctamente clasificadas por todos los modelos

En la Figura 5.14 se presentan ejemplos reales de imágenes fundus y OCT correspondientes a las clases Healthy y Glaucoma. Se puede observar que tanto las imágenes sanas como las asociadas a glaucoma presentan señales visuales claras. En el caso de *Healthy*, la retina muestra una estructura uniforme y sin alteraciones visibles, mientras que en las imágenes de *Glaucoma* es fácilmente identificable la depresión del disco óptico y la excavación aumentada, signos clínicos característicos de esta enfermedad. Estas diferencias morfológicas permiten que los modelos distingan estas clases con mayor fiabilidad, tal como reflejan las métricas obtenidas.

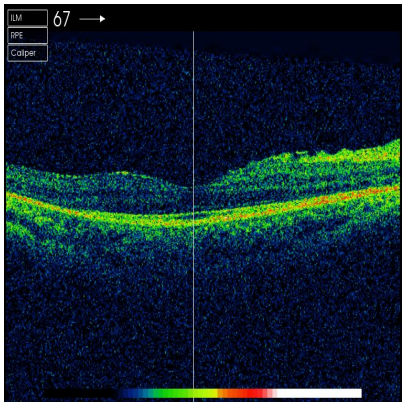
A continuación, se muestran imágenes clasificadas correctamente por la mayoría de los modelos, aunque no por todos. Estos ejemplos pertenecen a clases que tienden a presentar mayor variabilidad o ambigüedad visual, lo que puede dificultar su identificación en ciertas configuraciones:



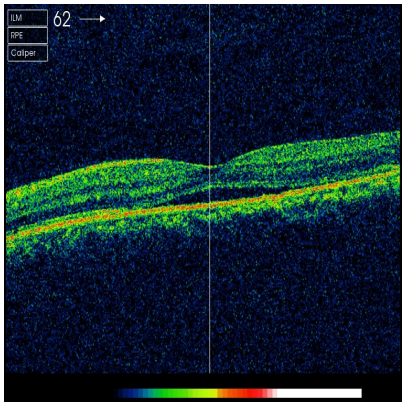
(a) Fundus DMAE



(b) Fundus CSR



(c) OCT DMAE



(d) OCT CSR

Figura 5.15: Ejemplos de imágenes correctamente clasificadas por la mayoría de modelos.

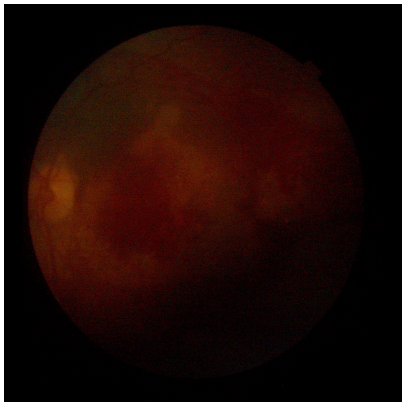
En la Figura 5.15 se incluyen ejemplos reales de las clases *DMAE* y *CSR*, que si bien han sido correctamente reconocidas por muchos modelos, han mostrado un rendimiento más inestable según el encoder y la técnica de fusión utilizada. Este comportamiento puede explicarse, por un lado, por la menor frecuencia de

estas clases en el conjunto de datos, lo cual limita el aprendizaje de patrones consistentes; y por otro, por la calidad variable de las imágenes.

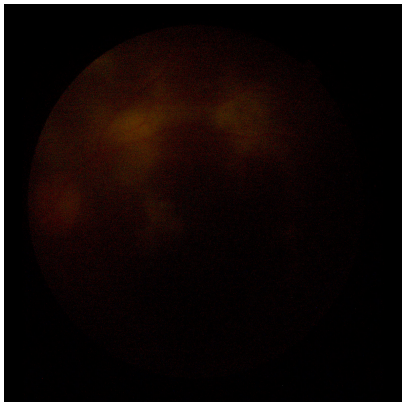
En las imágenes fundus (figuras 5.15a y 5.15b), se observan ciertas limitaciones como desenfoque o bajo contraste, que dificultan la identificación de señales relevantes. En el caso de *DMAE*, por ejemplo, no se aprecian claramente las drusas, mientras que en *CSR*, la retina muestra una apariencia casi normal.

Las imágenes OCT (figuras 5.15c y 5.15d) muestran señales algo más claras: en *DMAE*, se detecta una leve alteración en la línea del epitelio pigmentario, y en *CSR*, una separación visible de las capas retinianas, indicativa de fluido subretiniano. No obstante, estas señales pueden ser sutiles o variar según el paciente, lo que complica su detección automática.

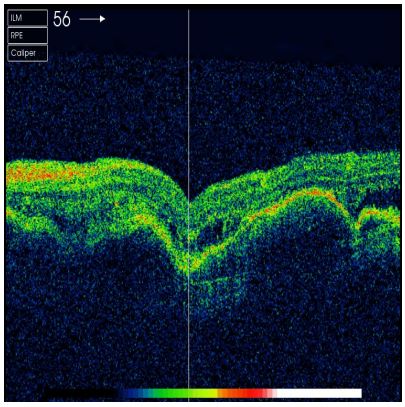
En línea con este análisis, a continuación se presentan imágenes que fueron clasificadas erróneamente por la mayoría de los modelos. Este tipo de casos suele estar asociado a una baja calidad de imagen, características poco representativas o patrones solapados con otras clases.



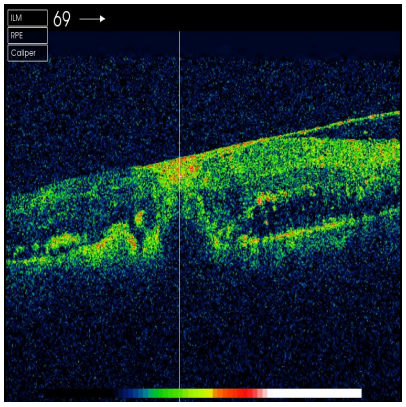
(a) Fundus DMAE



(b) Fundus ME



(c) OCT DMAE



(d) OCT ME

Figura 5.16: Ejemplos de imágenes correctamente clasificadas por la mayoría de modelos.

Tal como se observa en la Figura 5.16, varios factores pueden influir en la dificultad de clasificación de estas imágenes. En primer lugar, la baja calidad de las imágenes fundus es evidente. En las figuras 5.16a y 5.16b, se aprecia un contraste muy bajo, desenfoque y una iluminación deficiente, lo que complica la identificación de elementos visuales clave, como el nervio óptico o los vasos de la retina. Esta falta de nitidez dificulta que los modelos puedan detectar las señales necesarias para distinguir correctamente entre enfermedades.

Por otro lado, aunque las imágenes OCT presentadas en las figuras 5.16c y 5.16d tienen una calidad algo mejor, los patrones que muestran no siempre son lo suficientemente claros como para diferenciar las clases. En el caso de *DMAE*, se pueden observar leves irregularidades en la parte inferior de la retina, mientras que en *Edema Macular* aparecen zonas con acumulación de fluido. Sin embargo, estas alteraciones pueden parecerse entre sí o ser poco visibles, lo que lleva a errores en la clasificación, sobre todo si la enfermedad no está muy avanzada.

Este tipo de casos refleja la importancia de contar con imágenes más claras y bien representadas en el conjunto de datos. También pone en evidencia la necesidad de seguir mejorando los modelos para que puedan adaptarse mejor a la variabilidad de las imágenes reales y ser más precisos, incluso en situaciones complejas.

Capítulo 6

Conclusiones y trabajo futuro

6.1. Conclusiones

En este Trabajo de Fin de Grado se ha desarrollado un sistema de análisis de imágenes médicas oftalmológicas basado en técnicas de visión por computador y aprendizaje profundo. El objetivo principal ha sido entrenar, validar e interpretar modelos que trabajen con imágenes fundus y OCT para detectar y caracterizar patologías oculares de alta prevalencia, como la degeneración macular asociada a la edad (DMAE), el glaucoma, el edema macular (ME) y la coriorretinopatía serosa central (CSR).

Durante el desarrollo del sistema se ha logrado implementar una arquitectura de fusión multimodal capaz de integrar la información proveniente de ambas modalidades de imagen. Esto ha permitido explorar diferentes configuraciones y estudiar el impacto de cada una de ellas en el rendimiento final del modelo. Además, se han analizado e interpretado las salidas del sistema con el apoyo de métricas clínicas relevantes, así como con visualizaciones y análisis cualitativo de los resultados obtenidos.

La comprensión del contenido clínico de las imágenes también ha sido un aspecto fundamental. Para ello, se han estudiado los patrones característicos que permiten identificar distintas alteraciones en ambas modalidades. En las imágenes fundus se han reconocido estructuras como la mácula, el nervio óptico, los vasos retinianos y signos clínicos como drusas, exudados o hemorragias. Por su parte, las imágenes OCT han permitido evaluar la integridad de las capas retinianas, detectar edema, visualizar fluido subretiniano o identificar membranas epirretinianas.

A pesar de las limitaciones del tamaño de datos y la complejidad de algunas clases, el sistema desarrollado ha mostrado resultados prometedores, demostrando la utilidad de combinar modalidades de imagen para mejorar la precisión diagnóstica. Este trabajo sienta las bases para seguir explorando enfoques de fusión en oftalmología computacional, abriendo el camino hacia herramientas de ayuda al diagnóstico más eficaces y accesibles.

6.2. Trabajo futuro

Este sistema puede beneficiarse de varias mejoras y líneas de ampliación de cara al futuro. Algunas propuestas incluyen:

- **Experimentación con los hiperparámetros del modelo:** Ajustar los hiperparámetros de los modelos podría llevar a mejoras significativas en el rendimiento del sistema. Esto incluye la optimización de parámetros como la tasa de aprendizaje o el tamaño del lote. Además, se podrían explorar técnicas avanzadas de ajuste de hiperparámetros, como el *grid search*, para encontrar las configuraciones óptimas.
- **Aumento del tamaño del conjunto de datos de entrenamiento:** La efectividad de los modelos podría mejorarse significativamente con un mayor volumen de datos de entrenamiento. Adicionalmente, también es importante asegurar que los datos estén balanceados en términos de las clases de enfermedades representadas. Esto ayudará a que el modelo no se sesgue hacia clases más representadas, como ha sucedido en este caso debido a la falta de datos de ciertas clases.
- **Evaluación clínica en entorno real:** Probar el sistema con imágenes reales de pacientes en colaboración con centros oftalmológicos permitiría validar su aplicabilidad, detectar posibles errores y ajustar el modelo a las condiciones del entorno clínico.

Capítulo 7

Análisis de impacto del trabajo

El desarrollo de este Trabajo de Fin de Grado, centrado en el análisis e interpretación de imágenes OCT y de fondo de ojo para el diagnóstico de patologías oculares como la degeneración macular asociada a la edad (DMAE), glaucoma, edema macular o coriorretinopatía serosa central (CSR), tiene el potencial de generar un impacto positivo en múltiples dimensiones. En este capítulo se analiza cómo la comprensión de estas tecnologías puede mejorar la precisión diagnóstica en el ámbito médico, contribuir a la equidad en el acceso a la salud visual y ofrecer herramientas útiles en la formación de profesionales sanitarios. Asimismo, se examina su alineación con los Objetivos de Desarrollo Sostenible (ODS) de la Agenda 2030.

7.1. Impacto general

En términos generales, este trabajo puede impactar positivamente en las siguientes áreas:

- **Ámbito médico:** La familiarización con el uso de imágenes OCT y fundus permite a los profesionales de la salud ocular detectar patologías retinianas en fases tempranas, mejorando así el pronóstico visual de los pacientes. Además, el conocimiento detallado de los patrones imagenológicos típicos de enfermedades como la DMAE o el glaucoma puede facilitar el diagnóstico diferencial y la toma de decisiones clínicas más acertadas, reduciendo errores diagnósticos y mejorando la calidad asistencial.
- **Social:** La difusión del conocimiento sobre el uso de estas tecnologías puede contribuir a disminuir la desigualdad en el acceso a diagnóstico oftalmológico, especialmente en zonas rurales o con menos recursos. Promover herramientas de formación y concienciación sobre enfermedades visuales puede empoderar tanto a profesionales como a pacientes, favoreciendo una atención más equitativa y precoz.
- **Económico:** Una detección más temprana de enfermedades como el glaucoma o el edema macular permite iniciar tratamientos que previenen la

progresión del daño visual, lo cual reduce los costes a largo plazo derivados de la discapacidad visual. Además, el uso eficiente de tecnologías no invasivas como la OCT puede optimizar el tiempo de consulta y disminuir la necesidad de pruebas más costesas o invasivas.

- **Educativo:** Este trabajo puede servir como herramienta educativa para estudiantes de medicina, óptica y optometría, facilitando la comprensión de imágenes oftalmológicas clave. La inclusión de ejemplos visuales y descripciones accesibles contribuye a una mejor formación en diagnóstico visual, un área esencial en la práctica clínica.

7.2. Objetivos de Desarrollo Sostenible

El contenido de este trabajo está alineado con varios ODS de la Agenda 2030, entre los cuales destacan:

- **Salud y Bienestar (ODS 3):** El fortalecimiento de las capacidades diagnósticas mediante OCT y fundus puede mejorar significativamente el manejo de patologías visuales crónicas, favoreciendo un diagnóstico más temprano y tratamientos más eficaces, lo cual repercute en una mayor calidad de vida.
- **Educación de Calidad (ODS 4):** Este trabajo puede utilizarse como recurso educativo en instituciones académicas, promoviendo la capacitación de futuros profesionales en el uso e interpretación de imágenes oftalmológicas, fortaleciendo así la calidad de la enseñanza médica y optométrica.
- **Reducción de las Desigualdades (ODS 10):** La aplicación del conocimiento y tecnologías de diagnóstico visual en zonas con menor acceso a especialistas contribuye a reducir la desigualdad en el acceso a la salud ocular, fomentando la equidad y justicia social en la atención médica.

Bibliografía

[1] W. H. Organization. «World Report on Vision». (2019), dirección: https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment.

[2] D. Huang y colleagues. «Optical Coherence Tomography: Technology and applications». (1991), dirección: https://ieeexplore.ieee.org/document/6802013.

[3] Y.-C. Tham y colleagues. «Global Prevalence of Glaucoma and Projections». (2014), dirección: https://pubmed.ncbi.nlm.nih.gov/24974815/.

[4] O. A. Karen Allison Deepkumar Patel. «Epidemiology of Glaucoma: The Past, Present, and Predictions for the Future». (2021), dirección: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7769798/.

[5] G. W. Katie A Lucy. «Structural and Functional Evaluations for the Early Detection of Glaucoma». (2016), dirección: https://pmc.ncbi.nlm.nih.gov/articles/PMC5464747/.

[6] W. Wong y colleagues. «Global Prevalence of Age-related Macular Degeneration and Disease Burden Projection for 2020 and 2040: A Systematic Review and Meta-analysis». (2014), dirección: https://pubmed.ncbi.nlm.nih.gov/25104651/.

[7] B. Nicholson y colleagues. «Central Serous Chorioretinopathy: Update on Pathophysiology and Treatment». (2014), dirección: https://pmc.ncbi.nlm.nih.gov/articles/PMC3574296/.

[8] J. Yau y colleagues. «Global Prevalence and Major Risk Factors of Diabetic Retinopathy». (2012), dirección: https://diabetesjournals.org/care/article/35/3/556/28568/Global-Prevalence-and-Major-Risk-Factors-of.

[9] M. Data. «A Composite Retinal Fundus and OCT Dataset along with Detailed Clinical Markings for Extracting Retinal Layers, Retinal Lesions and Screening Macular and Glaucomatous Disorders». (2021), dirección: https://data.mendeley.com/datasets/trghs22fpg/3.

[10] Y. H. Madhurima Chaudhuri y colleagues. «Age-Related Macular Degeneration: An Exponentially Emerging Imminent Threat of Visual Impairment and Irreversible Blindness». (2023), dirección: https://pmc.ncbi.nlm.nih.gov/articles/PMC10300666/.

[11] S. Natarajan. «Advances in technology helps in early detection of vision disorders». (2013), dirección: https://pmc.ncbi.nlm.nih.gov/articles/PMC3917384/.

[12] M. P. McBee, O. A. Awan, A. T. Colucci et al. «Deep Learning in Radiology». (2018), dirección: https://www.academicradiology.org/article/S1076-6332(18)30104-1/fulltext.

[13] P. e. a. Yifan. «DeepSeeNet: A Deep Learning Model for Automated Classification of Patient-based Age-related Macular Degeneration Severity from Color Fundus Photographs». (2018), dirección: https://pubmed.ncbi.nlm.nih.gov/30471319/.

[14] National Center for Biotechnology Information (NCBI). «AREDS Dataset». (2024), dirección: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000001.v3.p1.

[15] Kaggle. «APTOS 2019 Blindness Detection». (2019), dirección: https://www.kaggle.com/competitions/aptos2019-blindness-detection.

[16] D. S. e. a. Kermany. «Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning». (2018), dirección: https://www.cell.com/cell/fulltext/S0092-8674(18)30154-5.

[17] J. e. a. De Fauw. «Clinically applicable deep learning for diagnosis and referral in retinal disease». (2018), dirección: https://www.nature.com/articles/s41591-018-0107-6.

[18] M. E. hospital. «DeepMind Updates». (2018), dirección: https://www.moorfields.nhs.uk/research/google-deepmind/google-deepmind-updates.

[19] «MultiEYE: Dataset and Benchmark for OCT-Enhanced Retinal Disease Recognition from Fundus Images». (2024), dirección: https://ar5iv.labs.arxiv.org/html/2412.09402v2.

[20] A. Ismail, M. A. Kassem, H. F. Hamed y M. Amin. «Explainable deep learning model for multimodal medical image fusion: A new benchmark and baseline». (2024), dirección: https://www.sciencedirect.com/science/article/pii/S0933365724000162.

[21] P. e. a. Rajpurkar. «CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning». (2017), dirección: https://arxiv.org/abs/1711.05225.

[22] A. e. a. Bustos. «PadChest: A large chest X-ray image dataset with multi-label annotated reports». (2020), dirección: https://arxiv.org/abs/1901.07441.

[23] H.-C. e. a. Shin. «Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning». (2016), dirección: https://pmc.ncbi.nlm.nih.gov/articles/PMC4890616/.

[24] S. e. a. Azizi. «Big Self-Supervised Models Advance Medical Image Classification». (2021), dirección: https://arxiv.org/abs/2101.05224.

[25] T. e. a. Schlegl. «Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery». (2017), dirección: https://arxiv.org/abs/1703.05921.

[26] H. N. D. et al. «A Multimodal Transfer Learning Approach Using PubMedCLIP for Medical Image Classification». (2024), dirección: https://ieeexplore.ieee.org/document/10531719.

[27] IBM. «Convolutional Neural Networks (CNNs)». (2023), dirección: https://www.ibm.com/think/topics/convolutional-neural-networks.

[28] P. C. Alsaleh A. «A space and time efficient convolutional neural network for age group estimation from facial images». (2023), dirección: https://peerj.com/articles/cs-1395/.

[29] A. Chaurasia. «Understanding Convolution Operations in CNN». (2020), dirección: https://medium.com/analytics-vidhya/understanding-convolution-operations-in-cnn-1914045816d4.

[30] H. Nelson, *Essential Math for AI*. O'Reilly Media, Inc., 2023.

[31] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan y M. Shah, «Transformers in vision: A survey», *ACM Computing Surveys (CSUR)*, vol. 54, n.º 10s, págs. 1-41, 2022.

[32] M. Notebook. «Convolutional Neural Networks - Basics». (2020), dirección: https://mlnotebook.github.io/post/CNN1/.

[33] A. e. a. Dosovitskiy. «An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale». (2020), dirección: https://arxiv.org/abs/2010.11929.

[34] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan y M. Shah, «Transformers in vision: A survey», *ACM Computing Surveys (CSUR)*, vol. 54, n.º 10s, págs. 1-41, 2022.

[35] A. e. a. Vaswani. «Attention is All You Need». (2017), dirección: https://arxiv.org/abs/1706.03762.

[36] Cohere. «What Are Transformer Models and How Do They Work?» (2024), dirección: https://cohere.com/blog/what-are-transformer-models.

[37] Cohere. «What Are Transformer Models?» (2024), dirección: https://cohere.com/llmu/what-are-transformer-models.

[38] Innovatiana. «Transformador de visión: fundamentos y aplicaciones». (2024), dirección: https://es.innovatiana.com/post/vision-transformer.

[39] Viso.ai. «Vision Transformers (ViT) in Image Recognition: Full Guide». (2023), dirección: https://viso.ai/deep-learning/vision-transformer-vit/.

[40] M. Oquab, T. Darcet, T. Moutakanni, H. Vo y et al., «DINOv2: Learning Robust Visual Features without Supervision», 2023. dirección: https://arxiv.org/abs/2304.07193.

[41] Encord. «DINOv2: Self-supervised Learning Model Explained». (2023), dirección: https://encord.com/blog/dinov2-self-supervised-learning-explained/.

[42] M. Oquab, T. Darcet, T. Moutakanni et al. «DINOv2: Learning Robust Visual Features without Supervision». (2023), dirección: https://arxiv.org/html/2304.07193v2.

[43] M. AI. «DINOv2: Exploring Self-Supervised Vision Transformers». (2023), dirección: https://blog.marvik.ai/2023/05/16/dinov2-exploring-self-supervised-vision-transformers/.

[44] Roboflow. «What is DINOv2? Self-Supervised Learning with Vision Transformers». (2023), dirección: https://blog.roboflow.com/what-is-dinov2/.

[45] «Brief Review — iBOT: Image BERT Pre-Training with Online Tokenizer». (2024), dirección: https://sh-tsang.medium.com/brief-review-ibot-image-bert-pre-training-with-online-tokenizer-85c32e47fee6.

[46] «Sinkhorn's theorem». (2024), dirección: https://en.wikipedia.org/wiki/Sinkhorn%27s_theorem.

[47] «Paper Review: DINOv2: Learning Robust Visual Features without Supervision». (2023), dirección: https://artgor.medium.com/paper-review-dinov2-learning-robust-visual-features-without-supervision-c56bd53b8e17.

[48] A. Radford, J. W. Kim, C. Hallacy et al., «Learning Transferable Visual Models From Natural Language Supervision», en *Proceedings of the 38th International Conference on Machine Learning (ICML)*, PMLR, 2021, págs. 8748-8763.

[49] OpenAI. «CLIP: Connecting text and images». (2021), dirección: https://openai.com/research/clip.

[50] Summergeometry. «A Deeper Understanding: OpenAI's CLIP Model». (2024), dirección: https://summergeometry.org/sgi2024/a-deeper-understanding-openais-clip-model/.

[51] R. Pulapakura. «Multimodal Models and Fusion: A Complete Guide». (2023), dirección: https://medium.com/@raj.pulapakura/multimodal-models-and-fusion-a-complete-guide-225ca91f6861.

[52] H. Tibebu. «Introduction to data fusion». (2023), dirección: https://medium.com/haileleol-tibebu/data-fusion-78e68e65b2d1.

[53] C. Qin, Z. Li, W. Zhao, X. He y J. Zhou, *MultiModality in Large Vision-Language Models: A Survey*, arXiv preprint, 2024. dirección: https://arxiv.org/html/2411.17040v1.

[54] R. Sinha, M. S. Hossain, G. Muhammad, M. Alsulaiman y W. Abdul. «Graphical representation for the early, intermediate and late fusions». (2021), dirección: https://www.researchgate.net/figure/Graphical-representation-for-the-early-intermediate-and-late-fusions-A-Early-fusion_fig2_352447987.

[55] M. Clinic. «Degeneración macular seca». (2024), dirección: https://www.mayoclinic.org/es/diseases-conditions/dry-macular-degeneration/symptoms-causes/syc-20350375.

[56] I. N. del Ojo (NEI). «Degeneración macular relacionada con la edad». (2021), dirección: https://www.nei.nih.gov/espanol/aprenda-sobre-la-salud-ocular/enfermedades-y-afecciones-de-los-ojos/degeneracion-macular-relacionada-con-la-edad.

[57] MedlinePlus. «Degeneración macular». (2023), dirección: https://medlineplus.gov/spanish/maculardegeneration.html.

[58] A. Cusumano. «Coriorretinopatía serosa central – Síntomas y diagnóstico». (2025), dirección: https://andreacusumano.com/es/retina-y-v%C3%ADtreo/Coriorretinopat%C3%ADa-serosa-central/.

[59] Wikipedia. «Coriorretinopatía serosa central». (2023), dirección: https://es.wikipedia.org/wiki/Coriorretinopat%C3%ADa_serosa_central.

[60] MedlinePlus. «Coroidopatía serosa central – Enciclopedia Médica». (2024), dirección: https://medlineplus.gov/spanish/ency/article/001612.htm.

[61] C. de Oftalmología Barraquer. «Edema macular: causas, síntomas y tratamientos». (2025), dirección: https://www.barraquer.com/patologia/edema-macular.

[62] I. O. Hoyos. «Retinopatía diabética: causas, síntomas y tratamientos». (2025), dirección: https://iohoyos.com/retinopatia-diabetica-causas-sintomas-y-tratamientos/.

[63] G. R. Foundation. «Datos y estadísticas sobre el glaucoma». (2023), dirección: https://glaucoma.org/es/articles/datos-y-estadisticas-sobre-el-glaucoma.

[64] Wikipedia. «Glaucoma». (2023), dirección: https://es.wikipedia.org/wiki/Glaucoma.

[65] N. E. Institute. «Glaucoma». (2024), dirección: https://www.nei.nih.gov/espanol/glaucoma.

[66] MedlinePlus. «Glaucoma – Enciclopedia Médica». (2024), dirección: https://medlineplus.gov/spanish/ency/article/001620.htm.

[67] F. Visión. «Retinografía – ¿Para qué sirve esta prueba?» (2025), dirección: https://futurovision.com/guia-del-paciente/retinografia/.

[68] G. research foundation. «Excavación del nervio óptico». (2023), dirección: https://glaucoma.org/es/articles/excavacion-del-nervio-optico.

[69] Miranza. «OCT (Tomografía de Coherencia Óptica)». (2022), dirección: https://miranza.es/diagnosticos/oct/.

[70] C. novovisión. «OCT – Tomografía de Coherencia Óptica – qué es y para qué sirve». (2019), dirección: https://www.clinicasnovovision.com/blog/oct-tomografia-de-coherencia-optica/.

[71] J. A. Clark y Contributors, *Pillow (PIL Fork) Documentation*. dirección: https://pypi.org/project/pillow/.

[72] «Albumentations». (), dirección: https://albumentations.ai/docs/.

[73] S. Project. «Contraste Adaptativo con Histograma Ecualizado (CLAHE)». (2024), dirección: https://siril.readthedocs.io/es/stable/processing/clahe.html.

[74] Baeldung. «GELU Activation Function – A Detailed Explanation». (2023), dirección: https://www.baeldung.com/cs/gelu-activation-function.

[75] García-Ferreira. «Conoce todo sobre el Dropout: qué es, cómo funciona y por qué usarlo». (2023), dirección: https://www.garcia-ferreira.es/conoce-todo-sobre-el-dropout/.

[76] P. with Code. «Layer Normalization - Method Overview». (2024), dirección: https://paperswithcode.com/method/layer-normalization.

[77] P. with Code. «AdamW - Method Overview». (2024), dirección: https://paperswithcode.com/method/adamw.

[78] P. with Code. «Weight Decay - Method Overview». (2024), dirección: https://paperswithcode.com/method/weight-decay.

[79] PyTorch. «torch.nn.CrossEntropyLoss». (2024), dirección: https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html.

[80] M. AI. «facebook/dinov2-base». (2024), dirección: https://huggingface.co/facebook/dinov2-base.

[81] OpenAI. «openai/clip-vit-base-patch32». (2024), dirección: https://huggingface.co/openai/clip-vit-base-patch32.

[82] «Performance Metrics Deep Dive». (2023), dirección: https://docs.ultralytics.com/guides/yolo-performance-metrics/#visual-outputs.

MARIO RUIZ VAQUETT

TFG_Mario_Ruiz_Vaquett_VFinal.pdf

Turnitin Memoria Final

TFG ETSIINF (Moodle PP)

Universidad Politecnica de Madrid

Detalles del documento

Identificador de la entrega

trn:oid:::1:3265136301

Fecha de entrega

31 may 2025, 2:25 p.m. GMT+2

Fecha de descarga

31 may 2025, 3:09 p.m. GMT+2

Nombre de archivo

31761_MARIO_RUIZ_VAQUETT_TFG_Mario_Ruiz_Vaquett_VFinal_83714_1205687829.pdf

Tamaño de archivo

13.0 MB

76 Páginas

21.862 Palabras

122.701 Caracteres

5% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Top Sources

0% Internet sources
0% Publications
5% Submitted works (Student Papers)

Top Sources

0% Internet sources
0% Publications
5% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1 Student papers
Universidad Politécnica de Madrid <1%

2 Student papers
Universidad Rey Juan Carlos <1%

3 Student papers
ucb <1%

4 Student papers
Universidad Católica San Pablo <1%

5 Student papers
Indiana University <1%

6 Student papers
Universidad de Salamanca <1%

7 Student papers
University of Melbourne <1%

8 Student papers
University of New South Wales <1%

9 Student papers
Universidad Catolica San Antonio de Murcia <1%

10 Student papers
University of Glasgow <1%

11 Student papers
UPAEP: Universidad Popular Autónoma del Estado de Puebla <1%

12 Student papers

Queen's University of Belfast <1%

13 Student papers

umb <1%

14 Student papers

Universidad de Valladolid <1%

15 Student papers

Reykjavík University <1%

16 Student papers

The University of Manchester <1%

17 Student papers

Universidad Nacional del Chimborazo <1%

18 Student papers

University of Westminster <1%

19 Student papers

Botswana International University of Science and Technology <1%

20 Student papers

University of Northumbria at Newcastle <1%

21 Student papers

Fundacion San Pablo Andalucia CEU <1%

22 Student papers

ICTS <1%

23 Student papers

UTEC Universidad de Ingeniería & Tecnología (NO TOCAR) <1%

24 Student papers

Centro Europeo de Postgrado - CEUPE <1%

25 Student papers

Universidad de Alicante <1%

26 Student papers

CORPORACIÓN UNIVERSITARIA IBEROAMERICANA <1%

27 Student papers

Infile <1%

28 Student papers

West Coast University <1%

29 Student papers

Babes-Bolyai University <1%

30 Student papers

Instituto Tecnológico y de Estudios Superiores de Occidente <1%

31 Student papers

Universidad de Alcalá <1%

32 Student papers

Universidad Internacional de la Rioja <1%

33 Student papers

University of Sydney <1%

34 Student papers

Universidad Autónoma Metropolitana-Xochimilco <1%

35 Student papers

Universidad TecMilenio <1%

36 Student papers

Universitat Politècnica de València <1%

37 Student papers

Florida A&M University <1%

38 Student papers

Middlesex University <1%

39 Student papers

Yeditepe University <1%

40 Student papers

National University of Ireland, Galway <1%

41 Student papers

University of Minnesota System <1%

42 Student papers

University of Sussex <1%

43 Student papers

Unviersidad de Granada <1%

44 Student papers

LUT-korkeakoulut <1%

45 Student papers

University of Northampton <1%

Este documento esta firmado por

	**Firmante**	CN=tfgm.fi.upm.es, OU=CCFI, O=ETS Ingenieros Informaticos - UPM, C=ES
	Fecha/Hora	Mon Jun 02 20:43:24 CEST 2025
	Emisor del Certificado	EMAILADDRESS=camanager@etsiinf.upm.es, CN=CA ETS Ingenieros Informaticos, O=ETS Ingenieros Informaticos - UPM, C=ES
	Numero de Serie	561
	Metodo	urn:adobe.com:Adobe.PPKLite:adbe.pkcs7.sha1 (Adobe Signature)