UNIVERSIDAD DE VALLADOLID

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE TELECOMUNICACIÓN

TRABAJO FIN DE MÁSTER

MÁSTER EN INGENIERÍA DE TELECOMUNICACIÓN

ANÁLISIS DE RETINOGRAFÍAS BASADO EN *DEEP LEARNING* PARA LA AYUDA AL DIAGNÓSTICO DE LA RETINOPATÍA DIABÉTICA

Autora:
Dña. Cristina Pinar Muñoz Zamarro

Tutora:
Dra. Dña. María García Gadañón

Valladolid, septiembre de 2022

TÍTULO:	**Análisis de retinografías basado en *Deep Learning* para la ayuda al diagnóstico de la Retinopatía Diabética**
AUTORA:	**Dña. Cristina Pinar Muñoz Zamarro**
TUTORA:	**Dra. Dña. María García Gadañón**
DEPARTAMENTO:	**Teoría de la Señal y Comunicaciones e Ingeniería Telemática**

TRIBUNAL

PRESIDENTE:	**Dr. D. Roberto Hornero Sánchez**
VOCAL:	**Dra. Dª. Miriam Antón Rodríguez**
SECRETARIO:	**Dr. D. Jesús Poza Crespo**
SUPLENTE 1:	**Dr. D. Salvador Dueñas Carazo**
SUPLENTE 2:	**Dr. D. Carlos Gómez Peña**
SUPLENTE 3:	**Dr. D. David González Ortega**

FECHA:	**septiembre de 2022**
CALIFICACIÓN:	

Agradecimientos

A todos los miembros del Grupo de Ingeniería Biomédica de la Universidad de Valladolid y en especial, a mi tutora María García Gadañón y a Roberto Romero Oraá, por apostar por mí una vez más y por su ayuda y comprensión en las diferentes etapas del trabajo.

A mis amigos y compañeros, por el gran apoyo que me han brindado siempre a lo largo del camino.

Por último, a mi familia por ser el pilar fundamental en los momentos más duros y por la confianza depositada en mí desde el primer día.

Muchas gracias a todos.

Resumen

La vista es uno de los sentidos más importantes para el ser humano. En los últimos años, el número de enfermedades que afectan a la visión ha aumentado considerablemente y se espera esta misma tendencia en los próximos años. Algunas de ellas, como por ejemplo la retinopatía diabética, el glaucoma o las cataratas, se han convertido en importantes causas de pérdida de visión a nivel mundial. Las alteraciones que provocan en el ojo humano se pueden observar en imágenes digitales, como las retinografías. Esta técnica es muy común y resulta muy útil para el diagnóstico de este tipo de patologías. La detección temprana es clave para evitar que la enfermedad llegue a sus estadios más avanzados y para que el tratamiento sea más efectivo, por lo que los pacientes deberían someterse a exámenes oftalmológicos frecuentes. No obstante, la creciente incidencia de estas enfermedades y la escasez de oftalmólogos especialistas provoca que el análisis manual de retinografías sea una tarea compleja y laboriosa. Es en este contexto donde los sistemas de cribado automático pueden resultar de gran utilidad para asistir a los oftalmólogos. A pesar de la gran efectividad de los sistemas basados en *Deep learning*, su aplicabilidad a la práctica clínica todavía no es muy evidente, como consecuencia de su carácter de "caja negra". Con el objetivo de solventar este problema, se ha desarrollado *Explainable Artificial Intelligence* (XAI), un conjunto de técnicas que tratan de explicar las decisiones que toman los modelos computacionales cuando se emplean para una tarea concreta.

El objetivo de este Trabajo Fin de Máster fue la aplicación de técnicas o métodos XAI para explicar los resultados obtenidos con un método automático de detección de patología ocular desarrollado en un estudio previo. Para ello, se ha hecho uso de dos bases de datos: una privada formada por un total de 1000 retinografías y otra pública conocida como RFMiD, que consta de 3200 retinografías. En ambas, las imágenes se han clasificado en dos clases correspondientes a fondos de ojo sanos y a fondos de ojo patológicos. Como clasificador, se ha empleado una red neuronal convolucional (CNN) profunda con arquitectura base DenseNet-121, junto con varias técnicas de optimización tales como *data augmentation*, *transfer learning*, *fine tuning* y *dropout*. Las técnicas XAI implementadas se corresponden con métodos de atribución y son las siguientes: *Shapley Additive Explanations* (SHAP), *Input x Gradient*, *Integrated Gradients* (IG), *SmoothGrad*, *DeepTaylor* y *Layer wise relevance propagation* (LRP).

En la base de datos privada, se ha conseguido alcanzar una precisión del 99.00%,

una sensibilidad del 100%, una especificidad igual al 98.00% y un AUC igual a 0.99. Los resultados obtenidos con la base de datos RFMiD son algo inferiores, pero también se supera el 90% de precisión. En concreto, se ha obtenido una precisión del 90.93%, una sensibilidad del 91.30%, una especificidad del 89.55% y un AUC igual a 0.97.

En cuanto a los métodos XAI, *Integrated Gradients* es el que obtuvo los mapas de atribución más comprensibles para el usuario, de ahí que sea considerado como el método óptimo. Concretamente, obtiene cada valor del mapa mediante el cálculo de una integral de gradientes y cuenta con un hiperparámetro que determina el número de pasos empleados para dicho cálculo. Los mejores resultados se obtuvieron con un valor de 160 pasos.

Los resultados obtenidos demuestran que los métodos XAI son efectivos para justificar las predicciones realizadas por las CNNs. La detección automática de patologías oculares mediante este tipo de redes supondría una importante ayuda para oftalmólogos especialistas ya que permitiría reducir su carga de trabajo, así como obtener diagnósticos más fiables. Además, la aplicación de XAI contribuiría a fomentar su aplicabilidad en entornos reales ya que los profesionales médicos podrían entender mejor los resultados obtenidos, lo que provocaría un aumento de su confianza en dichos sistemas.

Palabras clave

Cribado automático; *deep learning*; *explainable artificial intelligence*; redes neuronales convolucionales; retinopatía diabética.

Abstract

Sight is one of the most important senses for human beings. In recent years, the number of eye diseases has increased considerably and the same trend is expected in the coming years. Some of them, such as diabetic retinopathy, glaucoma or cataracts, have become major causes of vision loss worldwide. The alterations they cause in the human eye can be seen using digital images, such as fundus images. This technique is very common and useful for the diagnosis of this type of pathologies. Early detection is key to prevent the disease from reaching its most advanced stages and to make treatment more effective. Therefore patients should undergo frequent ophthalmological examinations. However, the increasing incidence of some diseases and the shortage of specialist ophthalmologists make the manual analysis of retinal images a complex and time-consuming task. In this context, automated screening systems can be very useful to assist ophthalmologists. Despite the great effectiveness of Deep learning-based systems, their application in clinical practice is still not very evident, as a consequence of their "black box" nature. In order to solve this problem, Explainable Artificial Intelligence (XAI) has been developed, a set of techniques that try to explain the decisions made by computational models when they are used for a specific task.

The aim of this study was the application of XAI techniques or methods to explain the results obtained with an automatic method for the ocular pathology detection developed in a previous study. To this end, two databases have been used: a private one consisting of 1000 fundus images and a public one, known as RFMiD, which consists of 3200 fundus images. In both datasets, the images have been classified into two classes corresponding to healthy fundus and pathological fundus images. As a classifier, a deep convolutional neural network (CNN) with DenseNet-121 base architecture has been used, together with several optimization techniques such as data augmentation, transfer learning, fine tuning and dropout. The implemented XAI techniques correspond to attribution methods and are the following: Shapley Additive Explanations (SHAP), Input x Gradient, Integrated Gradients (IG), SmoothGrad, DeepTaylor and Layer wise relevance propagation (LRP).

In the private database, the results achieved an accuracy of 99.00%, a sensitivity of 100%, a specificity of 98.00%, and an AUC equal to 0.99. Although the results with the RFMiD database were a little bit lower, they also exceeded 90% accuracy. Specifically, an accuracy of 90.93%, a sensitivity of 91.30%, a specificity of 89.55%

and an AUC equal to 0.97 were obtained.

As for the XAI methods, Integrated Gradients is the one that obtained the most user-friendly attribution maps, hence it is considered the optimal method. Specifically, it calculates each map value using a gradient integral and has a hyperparameter that determines the number of steps used. The best results were obtained with a value of 160 steps.

The results show that XAI methods are effective to justify the predictions made by CNNs. The automatic detection of ocular pathologies using this type of neural network would be an important aid for specialist ophthalmologists as it would reduce their workload, while obtaining more reliable diagnoses. Furthermore, the application of XAI would help to promote its applicability in real-world environments, as medical professionals would better understand the results obtained, thus increasing their confidence in these systems.

Keywords

Automatic screening; *deep learning*; *explainable artificial intelligence*; convolutional neural networks; diabetic retinopathy.

Glosario de acrónimos

ANN. Red neuronal artificial (*Artificial neural network*)

AUC. Área bajo la curva (*Area Under Curve*)

BD. Base de datos

CWs: Exudados algodonosos (*Cotton-Wool spots*)

DCNN. Red neuronal profunda (*Deep neural network*)

DIP. Diabetes en el embarazo

DL. Deep Learning

DM. Diabetes mellitus

DMAE. Degeneración macular asociada a la edad

DMG. Diabetes mellitus gestacional

DO. Disco óptico

DR. Desprendimiento de retina

ELM. Membrana limitante externa (*External Limiting Membrane*)

EMCS. Edema macular clínicamente significativo

EMD. Edema macular diabético

EOD. Enfermedad ocular diabética

EXs. Exudados duros

FFNN. Red neuronal prealimentada (*Feedforward Neural Network*).

FOV. Campo de visión (*Field Of View*)

GAA. Glaucoma de ángulo abierto

GAC. Glaucoma de ángulo cerrado

GCL. Capa de células ganglionares (*Ganglion Cell Layer*)

GIB. Grupo de Ingeniería Biomédica

GNT. Glaucoma de tensión normal

GPAA. Glaucoma primario de ángulo abierto

GSAA. Glaucoma secundario de ángulo abierto

GTN. Glaucoma de tensión normal

HEs. Hemorragias

HIP. Hiperglucemia en el embarazo

IA. Inteligencia artificial

ICG. Indocianina verde (*Indocyanine Green*)

IG. Integrated Gradients

ILM. Membrana limitante interna (*Inner Limiting Membrane*)

INL. Capa nuclear interna (*Inner nuclear layer*)

IOBA. Instituto Universitario de Oftalmobiología Aplicada

IPL. Capa plexiforme interna (*Inner Plexiform Layer*)

IRMAs. Anomalías microvasculares intrarretinianas (*Intraretinal Microvascular Abnormalities*)

LRP. Layer wise relevance propagation

MAs. Microaneurismas

ML. Machine Learning

MLP. Perceptrón multicapa (*Multilayer Perceptron*)

NFL. Capa de fibras nerviosas (*Nerve Fiber Layer*)

OACR. Oclusión de la arteria central retiniana

OAHR. Oclusión de la arteria hemirretiniana

OAR. Oclusión arterial retinana

OCT. Tomografía de coherencia óptica (*Optical Coherence Tomography*)

OCTA. Angiografía por tomografía de coherencia óptica (*Optical Coherence Tomography Angiography*)

OMS. Organización mundial de la salud

ONH. Cabeza del nervio óptico

ONL. Capa nuclear externa (*Outer Nuclear Layer*)

OPL. Capa plexiforme externa (*Outer Plexiform Layer*)

ORAO. Oclusión de rama de la arteria retiniana

ORVR. Oclusión de la rama venosa retiniana

OVCR. Oclusión de la vena central de la retina

OVHR. Oclusión de la vena hemirretiniana

OVR. Oclusión venosa retiniana

PIO. Presión intraocular

PPA. Atrofia peripapilar (*Peripapillary Atrophy*)

RDNP. Retinopatía diabética no proliferativa

RDP. Retinopatía diabética proliferativa

ROC. Característica de operación del receptor (*Receiver Operating Characteristic*)

ROI. Región de interés

RPE. Epitelio pigmentado de la retina (*Retinal Pigment Epithelium*)

SGD. Descenso de gradiente estocástico (*Stochastic Gradient Descent)*

SHAP. SHapley Additive exPlanations

SLO. Oftalmoscopia de láser de barrido (*Scanning Laser Ophthalmoscope*)

SVM. Máquina de vectores soporte (*Support Vector Machine*)

UVa. Universidad de Valladolid

VEGF. Factor de crecimiento endotelial vascular

XAI. Inteligencia artificial explicable (*eXplainable Artificial Intelligence*)

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Introducción

En este capítulo se introducen aquellas ideas o conceptos básicos que se van a utilizar a lo largo del trabajo. Primero se expone brevemente la fisiopatología de la Diabetes Mellitus (DM) y la Retinopatía Diabética (RD), una complicación visual derivada de la misma. También se incluyen otras enfermedades retinianas que provocan alteraciones en la vista y que pueden diagnosticarse empleando imágenes oftalmológicas. Posteriormente, se explica la utilidad de diferentes modalidades de imagen médica en el diagnóstico de patologías oculares. Se hace especial hincapié en las retinografías o imágenes de fondo de ojo y en cómo se puede aplicar *Deep Learning* (DL) para la ayuda al diagnóstico de enfermedades oculares a través de ellas. Posteriormente, se introduce el concepto de inteligencia artificial explicable (*eXplainable Artificial Intelligence*, XAI) en el contexto de la ayuda al diagnóstico de enfermedades oculares, lo que constituye la principal aportación de este TFM. Por último, se recogen las hipótesis del trabajo, así como los objetivos y la metodología seguida en el TFM, así como la estructura que se ha seguido en esta memoria.

1.2. Diabetes Mellitus

La DM es un trastorno metabólico crónico caracterizado por niveles elevados de glucosa en sangre (hiperglucemia). Se debe a defectos en la secreción de insulina, lo que provoca a su vez alteraciones en el metabolismo de proteínas, carbohidratos y grasas (Kharroubi & Darwish, 2015). Algunos de los síntomas más frecuentes de la DM son el aumento de la sed, poliuria, fatiga, visión borrosa, pérdida de peso inexplicable y entumecimiento en los pies o en las manos (National Institutes of Health, 2016). Algunos pacientes diabéticos son asintomáticos durante los primeros años de la enfermedad. No obstante, la gravedad de los síntomas depende de la duración y del tipo de DM. La DM no controlada puede provocar enfermedades cardiovasculares graves, nefropatías (enfermedades renales), neuropatías (lesiones

en los nervios), complicaciones en el embarazo que pueden desembocar en daños en órganos del feto y complicaciones visuales, tales como la RD, que pueden provocar la pérdida de visión o incluso, en algunos casos, ceguera (International Diabetes Federation, 2020).

En la actualidad, se estima que, a nivel mundial, el número de personas (en el rango de edad de 20 a 79 años) que padecen esta enfermedad está en torno a los 537 millones, lo que equivale prácticamente a un 10% de la población. No obstante, las estimaciones en la prevalencia mundial apuntan a un aumento en los próximos años, llegando a 643 millones (11.3%) en 2030 y a 783 millones (12.2%) en 2045 (International Diabetes Federation, 2021). Además, tal y como se puede observar en la Figura 1.1, en países de ingresos medios y bajos (como África) el número de pacientes diabéticos aumenta considerablemente, pues se estima que 3 de cada 4 adultos con DM viven en estos países (International Diabetes Federation, 2021).

Dada su gran incidencia, en 2021 esta enfermedad fue responsable de un gasto sanitario mundial estimado en 966.000 millones de dólares. Aun así, excluyendo los riesgos de mortalidad asociados a la COVID-19, se estima que 6.7 millones de adultos aproximadamente habrían muerto a causa de DM o de las complicaciones que provoca (International Diabetes Federation, 2021). Por lo tanto, es de vital importancia llevar controles periódicos para poder detectar dicha enfermedad a tiempo.

La DM se puede clasificar en diferentes tipos que dependen de la gravedad de la enfermedad. No obstante, esta clasificación a veces no es sencilla ya que muchos pacientes (sobre todo en adultos jóvenes) no encajan únicamente en un solo tipo (Kharroubi & Darwish, 2015). Actualmente se pueden distinguir tres tipos diferentes de DM (International Diabetes Federation, 2021):

- **Diabetes tipo I (DM insulinodependiente o juvenil).** La causa de la enfermedad se debe a un proceso autoinmune en el que el sistema inmunitario del organismo ataca a las células β del páncreas encargadas de producir insulina. Como consecuencia de ello, el cuerpo produce muy poca cantidad o nada de insulina. Aunque este tipo de DM puede desarrollarse a cualquier edad, suele diagnosticarse con mayor frecuencia en niños o en adultos jóvenes. Las personas que padecen diabetes tipo 1, requieren inyecciones diarias de insulina para mantener en un rango adecuado los niveles de glucosa en sangre.

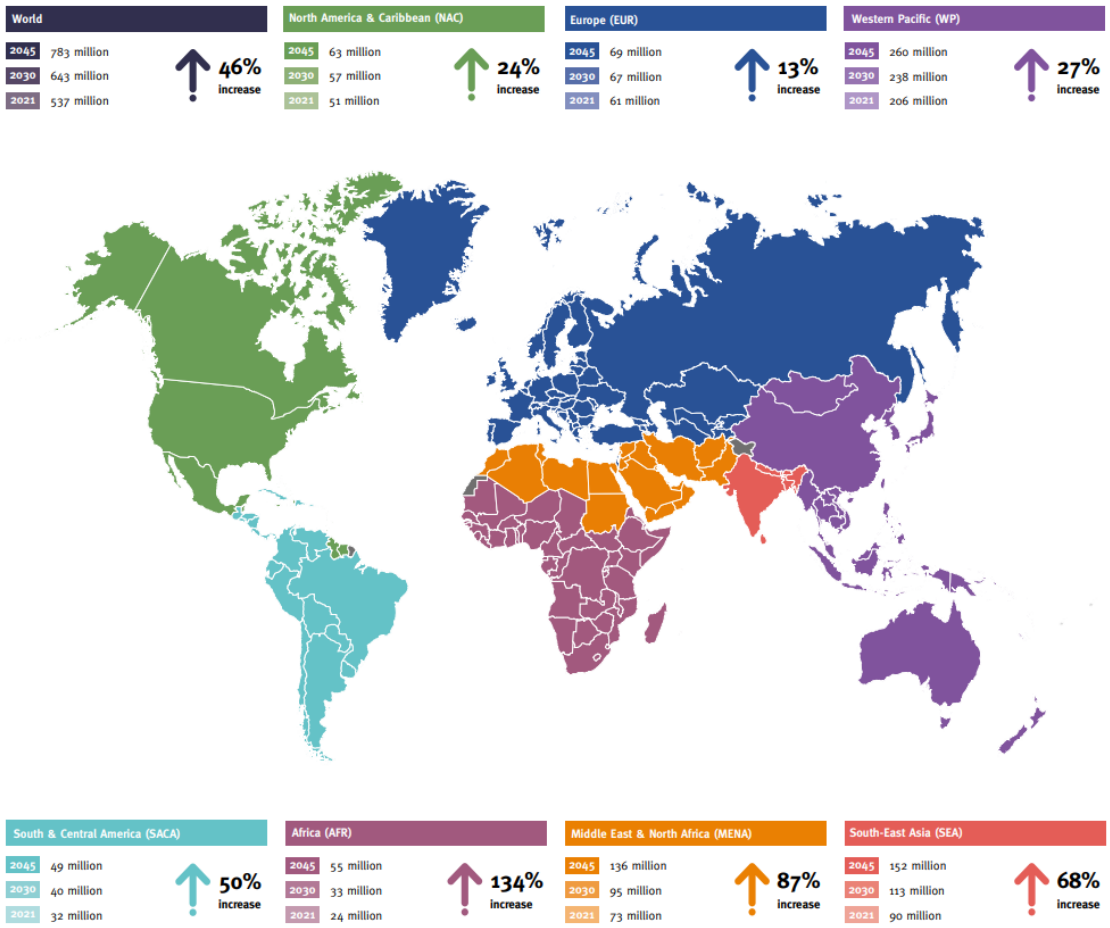


Figura 1.1. Número de personas de 20-79 años con diabetes a nivel mundial y por regiones, estimado de 2021-2045 (International Diabetes Federation, 2021).

- **Diabetes tipo II (DM no insulinodependiente o adulta).** Es el tipo de DM más común ya que alrededor del 90% de la población mundial la padece. En ella, la hiperglucemia se debe a la incapacidad de las células del organismo para responder plenamente a la insulina. Esto se conoce como resistencia a la insulina. Como consecuencia de la DM, la hormona de la insulina es menos eficaz, lo que conlleva a un aumento en su producción. Este incremento inadecuado de insulina puede desarrollarse con el tiempo debido a que las células β son incapaces de atender la demanda. Los síntomas son similares a la DM de tipo 1 aunque, en general, son de menor gravedad. Incluso, se puede dar la posibilidad de que la enfermedad sea completamente asintomática. Esto provoca que a casi la mitad de las personas con diabetes de tipo 2 no se les diagnostica la enfermedad a tiempo, lo que provoca la aparición de otras complicaciones tales como la RD, enfermedades cardiacas, cerebrovasculares, etc. En cuanto al tratamiento, es de vital importancia llevar un estilo de vida que incluya una dieta saludable, practicar actividad física de manera regular y no

fumar. No obstante, en algunos casos es necesario también la medicación oral con el objetivo de controlar los niveles de glucosa en sangre.

- **Hiperglucemia en el embarazo (HIP).** La HIP se puede clasificar a su vez como diabetes pregestacional, DM gestacional (DMG) o diabetes en el embarazo (DIP). El primer tipo incluye a aquellas mujeres con diabetes tipo 1 o tipo 2 que fueron diagnosticadas antes del embarazo. La DMG puede aparecer en cualquier momento dentro del periodo prenatal y no suele persistir tras el parto. Según las últimas estimaciones, entre un 75% y un 90% de los casos de HIP son de tipo DMG. Finalmente, la DIP se aplica a aquellas mujeres embarazadas que padecen hiperglucemia diagnosticada por primera vez durante el periodo de gestación y que cumple con los criterios de la Organización Mundial de la Salud (OMS) de diabetes en el estado de no embarazo. Los síntomas manifiestos de HIP son poco frecuentes y suelen ser complejos de diferenciar de los síntomas normales del embarazo. Por esta razón, es muy recomendable la realización de pruebas entre las semanas 24 y 28 que permitan el cribado de la enfermedad, e incluso antes del embarazo para las mujeres consideradas de alto riesgo.

1.3. Retinopatía diabética

La RD es una afección causada por la DM que puede provocar un deterioro en la visión ya que afecta en gran medida a los vasos sanguíneos de la retina y disminuye el revestimiento interno sensible a la luz del fondo de ojo (Nagpal et al., 2021). De hecho, es considerada una de las principales causas de ceguera evitable en la población adulta activa. Para evitar la pérdida de visión es clave la detección en sus primeras etapas para que el tratamiento sea más efectivo y se evite el avance de la enfermedad. No obstante, la detección temprana es difícil puesto que la enfermedad es asintomática hasta sus estadios avanzados (Beaser et al., 2018). Por esta razón, existen estudios que demuestran que la prevalencia mundial de ceguera derivada de esta enfermedad ocular se ha incrementado de un 14.9% a un 18.5% en el rango de tiempo comprendido entre 1990 y 2020 (Teo et al., 2021).

Con la progresión periódica de la RD, aparecen lesiones y anomalías en el fondo del ojo cuya detección permite identificar diferentes fases de la enfermedad. Entre las lesiones retinianas más comunes que se forman a partir de la rotura de los vasos sanguíneos de la retina, se encuentran: microaneurismas (MAs), hemorragias (HEs), exudados duros (EXs) y exudados algodonosos (CWs). A continuación, se explica con mayor detalle cada una de ellas (D. Das et al., 2022; Triwijoyo et al., 2020).

- **Microaneurismas.** Dilataciones capilares localizadas, de color rojo y con estructura sacular. Pueden aparecer en grupos o de forma aislada y su número y extensión depende de la fase en la que se encuentre la enfermedad. No obstante, tienen dimensiones típicas de entre 10 y 100 μm, de ahí que aparezcan como pequeñas manchas redondas en las imágenes de fondo de ojo.

- **Hemorragias.** Distorsiones estructurales en las paredes de los vasos sanguíneos de la retina que provocan fugas de sangre a través de los mismos. Poseen formas muy diferentes, que van desde puntos rojos redondos con márgenes nítidos (cuando aparecen en zonas intermedias de la retina) hasta puntos más indefinidos en forma de fuego (cuando se localizan próximas a los vasos principales). Debido a su similitud con las MAs, se las suele agrupar a ambas con el nombre de lesiones rojizas.

- **Exudados duros.** Pequeños depósitos intrarretinianos brillantes de color blanco amarillento con límites bien definidos que contienen proteínas y lípidos extracelulares como consecuencia de la fuga de sangre de los capilares retinianos anormales. En cuanto a su forma, pueden presentarse como pequeñas motas hasta grandes manchas e incluso evolucionar de manera gradual hasta convertirse en construcciones anulares denominadas circinadas.

- **Exudados algonodosos.** Manchas turbias irregulares con límites difusos o poco definidos y de mayor tamaño que los EXs. Suelen ser de color blanquecino o grisáceo y aparecen como consecuencia de la isquemia local que conduce a la interrupción del flujo axoplásmico.

En base a la aparición de las lesiones retinianas anteriormente comentadas, la RD se puede clasificar en diferentes tipos, de acuerdo con los grados de severidad de la enfermedad. Con el objetivo de simplificar su clasificación, en 2003 se creó la escala internacional de gravedad de la enfermedad clínica de la RD (*International Clinical Disease Severity Scale for DR*) (Wilkinson et al., 2003). Esta escala se basa en los resultados de dos estudios, el estudio epidemiológico de la RD de Wisconsin (WESDR) y el estudio sobre el tratamiento temprano de la RD (ETDRS).

Concretamente, se diferencian cinco grados de severidad, que a su vez se agrupan en dos grandes tipos como son: la RD no proliferativa (RDNP), asociada a las primeras fases y la RD proliferativa (RDP), asociada a las fases más avanzadas. A continuación, se explica en orden cada uno de ellos (Wu et al., 2013).

- **Grado 0. Sin retinopatía aparente.** Es el primer grado considerado en la escala y hace referencia a aquellos casos en los que no existe ninguna lesión ni cambio relevante en el fondo de ojo del paciente diabético.

- **Grado 1. RDNP leve.** Se considera la etapa más temprana de la enfermedad y se caracteriza por la presencia de algunos MAs en la retina.

- **Grado 2. RDNP moderada.** Se trata del segundo estadio de la enfermedad y en él, se observa la aparición de más lesiones aparte de MAs, tales como HEs intrarretinianas. Además, en algunos casos también se detectan EXs y CWs junto con dilataciones venosas.

- **Grado 3. RDNP severa.** Tercer grado de la enfermedad caracterizado por la presencia de cualquier lesión retiniana ya comentada. También, en esta etapa, se pueden observar anomalías microvasculares intrarretinianas (IRMAs). Éstas consisten en vasos finos tortuosos que se forman en la retina, debido a la fuga de la hormona del factor de crecimiento endotelial vascular (VEGF) por la obstrucción de los vasos sanguíneos (Patil & Daigavane, 2020). Los datos del ETDRS han demostrado que, aquellos pacientes que padecen DM tipo 2 y alcanzan este grado de severidad, tienen un 50% de posibilidades de desarrollar características de alto riesgo para la pérdida de visión, si no se realiza el tratamiento con láser (F. Ferris, 1996).

- **Grado 4. RDP.** Es la etapa más avanzada de la enfermedad y se caracteriza por la neovascularización (aparición de nuevos vasos sanguíneos) de las principales estructuras del ojo tales como el disco óptico (DO), la retina y el iris. Además, este grado puede llegar a provocar la aparición de una hemorragia vítrea o incluso, un desprendimiento de retina (DR).

En la Figura 1.2 se muestra un ejemplo de imagen de fondo de ojo para cada grado de severidad de la RD anteriormente comentado.

1.4. Otras enfermedades de la retina

Al grupo de afecciones oculares que están ligadas con la DM se las conoce como enfermedad ocular diabética (EOD) y comprende las siguientes: la RD, el edema macular diabético (EMD), el glaucoma y las cataratas (National Eye Institute, 2017). El desarrollo de la EOD en estado grave comienza con la aparición de vasos

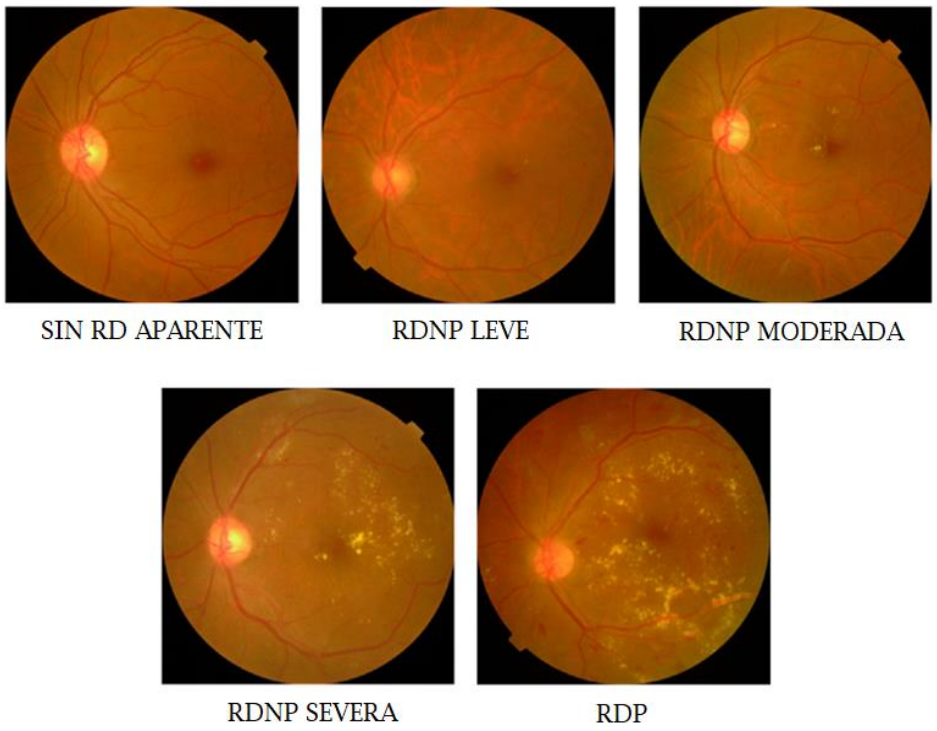


Figura 1.2. Ejemplos de imágenes de fondo de ojo para cada grado de severidad de RD (adaptada de Sikder et al., 2021).

sanguíneos irregulares, daños en el nervio óptico y la formación de EXs en la región de la mácula. No obstante, existen también otras enfermedades oculares que aparecen en la retina y que pueden provocar visión borrosa irreversible, metamorfopsia, defecto del campo visual o incluso ceguera si no se detectan a tiempo. Algunas de estas últimas son la degeneración macular asociada a la edad (DMAE), la oclusión venosa retiniana (OVR), la oclusión arterial retiniana (OAR) y el DR (Cen et al., 2021). Por lo tanto, la mejor defensa para todas las enfermedades oculares son los controles o exámenes regulares ya que muchas de ellas no siempre presentan síntomas.

A continuación, se explica brevemente en qué consisten y cómo afectan a la retina las distintas enfermedades mencionadas, a excepción de la RD puesto que ya se ha realizado una explicación detallada de la misma en el apartado anterior.

1.4.1. Edema macular diabético

El EMD consiste en un engrosamiento de la retina causado por la acumulación de líquido intrarretiniano tanto en la capa plexiforme interna como en la externa. Para caracterizar su gravedad y su consiguiente tratamiento, se utiliza el término edema

macular clínicamente significativo (EMCS), el cual fue definido por el ETDRS (Early Treatment Diabetic Retinopathy Study Research Group, 1991). El EMD es considerado EMCS cuando se cumple alguna de las siguientes condiciones: engrosamiento de la retina dentro de los 500 μm del centro de la mácula (una de las principales estructuras oculares de la retina que se explicará con mayor detalle posteriormente), presencia de EXs dentro de los 500 μm del centro de la mácula si están asociados con el engrosamiento de la retina adyacente y, por último, una zona o zonas de engrosamiento de la retina de 1 área de disco (Musat et al., 2015; Tripathy et al., 2015).

En la mayoría de los casos, el EMD tiende a ser una enfermedad crónica y, al igual que con otras enfermedades oculares, la detección temprana es vital para evitar la pérdida visual. Aunque existen casos de recuperación espontánea, se calcula que el 24% de casos con EMCS y el 33% con EMCS y afectación central, podrían perder moderadamente su visión si no se tratan en un plazo de tres años (Ferris et al., 1987). Además, muchos estudios han investigado el efecto de diferentes condiciones en la incidencia del EMD, tales como la microalbuminuria, la presión arterial diastólica, la nefropatía diabética y el elevado nivel de lípidos. Por lo tanto, se recomienda que todos los pacientes con EMD intenten normalizar dichas condiciones dentro de sus rangos de valores adecuados (Musat et al., 2015). Finalmente, en cuanto al tratamiento del EMD, recientes y rigurosos ensayos clínicos han demostrado que la fotocoagulación con láser ya no es efectiva para su tratamiento, pasando a ser terapia de primera línea el tratamiento del VEGF. Los esteroides mantienen su papel en el tratamiento del EMD, cuando éste es diagnosticado como crónico persistente. No obstante, los cambios de paradigma en la terapia van acompañados de los avances sustanciales en el diagnóstico (Schmidt-Erfurth et al., 2017).

1.4.2. Glaucoma

La enfermedad ocular conocida como glaucoma comprende a un grupo de neuropatías ópticas progresivas que se caracterizan por una degeneración de las células ganglionares y de las capas de fibras nerviosas de la retina, lo que provoca modificaciones en la cabeza del nervio óptico (ONH). Estos daños en el nervio óptico están relacionados con la presión intraocular (PIO) y son los que provocan la perdida de células ganglionares de la retina (Allison et al., 2020). Se pueden diferenciar dos tipos principales de glaucoma, el primario y el secundario, los cuales a su vez se dividen en dos subtipos según la anatomía y la fisiopatología adyacente, que son glaucoma de ángulo abierto (GAA) y de ángulo cerrado (GAC). Mientras que el glaucoma primario (o idiopático) es la consecuencia de la presencia de un GAA y GAC

sin causa identificable, el glaucoma secundario se origina debido al aumento de la PIO. El GAA puede a su vez clasificarse en tres subtipos que son glaucoma primario de ángulo abierto (GPAA), glaucoma de tensión normal (GTN) y glaucoma secundario de ángulo abierto (GSAA). Las diferencias que existen entre estos tres últimos son las siguientes: el GPAA se debe a un aumento de la PIO con progresión del nervio óptico, el GNA se caracteriza por una PIO normal con progresión y neuropatía óptica y, finalmente, en el GSAA la PIO es elevada y existe también neuropatía óptica (Harasymowycz et al., 2016).

La identificación de los factores de riesgo es vital para evitar el desarrollo del glaucoma. Entre los más importantes destaca la edad, puesto que el riesgo de padecer la enfermedad aumenta con los años y el género. Respecto a este último, ciertos estudios han demostrado que el género masculino es un factor predictivo útil para la aparición del GPAA, mientras que las mujeres tienen un mayor riesgo de padecer GAC. Otros factores también considerados de riesgo son la genética e historial familiar, es decir, si existen antecedentes familiares y la raza. Se estima que la prevalencia del GPAA en la población negra estadounidense es seis veces mayor respecto a la población blanca (McMonnies, 2017).

Si no se trata, el glaucoma puede acelerar la pérdida de visión permanente o la ceguera. Los tratamientos para esta enfermedad dependen del tipo y gravedad de la misma. Entre ellos, se incluyen colirios que ayudan a aumentar el drenaje para aliviar la PIO, el tratamiento con láser (trabeculoplastia), la cirugía convencional, que es más invasiva pero también suele conseguir mejores resultados, o bien una combinación de todos los anteriores (National Eye Institute, 2022a).

1.4.3. Cataratas

A nivel mundial, la enfermedad ocular conocida como cataratas afecta a unos 18 millones de personas y la frecuencia es entre 2 y 5 veces mayor en pacientes con DM (Javadi & Zarei-Ghanavati, 2008). Las cataratas aparecen cuando el cristalino se nubla afectando a la visión. El cristalino es la parte del ojo que ayuda a enfocar la luz o las imágenes sobre la retina. Éste está compuesto principalmente por agua y por proteínas que, con la edad, pueden acumularse provocando que una pequeña parte del cristalino se nuble, dando lugar a lo que se conoce por catarata. En pacientes sanos, la luz pasa a través del cristalino transparente hacia la retina y en ella se transforma en señales nerviosas que se envían al cerebro. Por lo tanto, si el cristalino es transparente, entonces la retina será capaz de recibir una imagen clara, mientras que, si existe una catarata, entonces la imagen se percibirá borrosa (National Eye Institute, 2022c).

Si bien las cataratas no se transmiten de un ojo al otro, sí que pueden aparecer en cualquiera de los dos o incluso en ambos. Existen diferentes tipos de cataratas dependiendo de la causa de su origen. Entre las más comunes se encuentra la catarata asociada a la edad, que, a medida que se envejece, se puede desarrollar como consecuencia de cambios naturales en el cristalino del ojo (Asbell et al., 2005). La catarata traumática es una secuela muy común de los traumatismos oculares contundentes o penetrantes que alteran las fibras del cristalino (American Academy of Ophthalmology, 2016). La catarata por radiación se origina por descargas eléctricas, radiaciones ionizantes (rayos X) o por la exposición a la energía infrarroja (Chodick et al., 2008). La catarata congénita o infantil es aquella que aparece en niños durante su primer año de vida aunque también pueden nacer con ella, tratándose, por tanto, de una enfermedad hereditaria (Bell et al., 2020). Finalmente, la catarata secundaria se origina después de otras cirugías como aquellas para corregir otras cataratas o bien otras patologías oculares como, por ejemplo, el glaucoma (Matsushima et al., 2008; National Eye Institute, 2022c).

Entre los síntomas más comunes de esta patología ocular se encuentran la visión borrosa u opaca, la visión doble o imágenes múltiples en un ojo, el destello, la mala visión nocturna y el desteñido de los colores, entre otros. Para tratar de hacer frente a esta enfermedad, se suele utilizar la corrección con gafas refractivas en las primeras fases. Si la catarata ya se encuentra en una etapa madura e interfiere en las actividades rutinarias, se puede aconsejar la cirugía, que es el único tratamiento eficaz y consiste en eliminar el cristalino opaco y reemplazarlo por una lente artificial (National Eye Institute, 2022c; Nizami & Gulani, 2019).

1.4.4. Degeneración macular asociada a la edad

La DMAE es una enfermedad crónica progresiva que afecta a la parte central de la retina (mácula) y es considerada también una de las principales causas de pérdida de visión en todo el mundo. No causa dolor y, en sus etapas más tempranas, los pacientes afectados apenas notan cambios en su visión. Es en las últimas fases de la enfermedad donde se produce la mayor parte de la pérdida visual debido a dos procesos: la DMAE neovascular (o también conocida como DMAE húmeda) y la atrofia geográfica (o DMAE seca tardía). En la primera de ellas, la neovascularización coroidea afecta a la retina neural, lo que provoca que los fluidos, los lípidos y la sangre se filtren dando lugar a cicatrices finas. Sin embargo, en la atrofia geográfica, se produce una atrofia progresiva del epitelio pigmentario de la retina, los fotorreceptores y la coriocapilaridad (Lim et al., 2012).

Entre los factores de riesgo de sufrir DMAE se encuentra el tabaquismo, la obesidad,

la exposición a la luz solar y las enfermedades cardiovasculares. Además, existen estudios que demuestran que las personas que sufren DMAE tienen también un mayor riesgo de padecer accidentes cerebrovasculares (Lim et al., 2012). En cuanto a los posibles tratamientos, conviene señalar que ayudan a ralentizar y evitar el avance de la enfermedad, pero en ningún caso sirven para erradicarla completamente. La DMAE neovascular puede tratarse mediante inyecciones en el ojo, terapia fotodinámica o con cirugía láser. Para el segundo tipo (DMAE seca) el grupo de investigación AREDS (*Age-Related Eye Disease*) del Instituto Nacional de la Salud (NIH) encontró que las dosis altas de zinc y antioxidantes ayudaban a reducir el riesgo de sufrir la enfermedad. No obstante, en la actualidad los oftalmólogos no lo prescriben a todos los pacientes con este tipo de DMAE, sino que se recomienda una dieta saludable que incluya ácidos grasos omega-3, así como luteína y zeaxantina. Además, en los últimos 10 años se han producido importantes mejoras en el tratamiento de la DMAE como por ejemplo el descubrimiento de la inyección intravítrea de un fármaco anti-VEGF, el cual abre una nueva perspectiva de medidas terapéuticas para esta enfermedad tan crítica (Jaffe et al., 2019; National Eye Institute, 2021).

1.4.5. Oclusión venosa retiniana

La enfermedad conocida como oclusión venosa retiniana (OVR) consiste en la obstrucción de alguna de las pequeñas venas que transportan la sangre desde la retina, lo que provoca la aparición de HEs y fugas de líquido de los vasos sanguíneos bloqueados. Es considerada la enfermedad vascular retiniana más común después de la RD y, dependiendo de cuál sea la zona de drenaje venoso ocluido, se pueden distinguir tres tipos diferentes: oclusión de la vena central de la retina (OVCR), oclusión de la vena hemirretiniana (OVHR) y oclusión de la rama venosa retiniana (ORVR) (Karia, 2010). En la OVCR se obstruye la vena principal de la retina posterior a la lámina cribosa del nervio óptico y suele estar causada por una trombosis (Blair & Czyz, 2021). La OVHR es una variante de la primera que afecta a la mitad superior o inferior de la retina y que se desarrolla debido a una variación anatómica en la ONH (Srivastava & Fekrat, 2006). Por último, la ORVR se debe a la obstrucción de una rama de la vena retiniana en un cruce arteriovenoso, lo que provoca daños en las células endoteliales y la formación de trombos (Cochran et al., 2018).

El síntoma más claro de la OVR en general es la pérdida visual variable e indolora junto con cualquier combinación de lesiones en el fondo de ojo tales como tortuosidad retiniana vascular, HEs (en forma de mancha y de llama), CWs, hinchazón del DO e incluso EMD. En una OVCR, las HEs se pueden observar en cualquiera de los cuatro cuadrantes del fondo del ojo, mientras que en una OVHR se

limitan exclusivamente al hemisferio superior o inferior. Finalmente, en una ORVR, las HEs se localizan principalmente en la zona drenada por la rama venosa de la retina ocluida (Karia, 2010). En cuanto a los tratamientos frente a esta enfermedad, el único que ha revelado efectos beneficiosos es la hemodilución, siempre y cuando se realice con prontitud tras el diagnóstico de la OVR. No obstante, existen muchas contraindicaciones tales como la DM, hipertensión no controlada, insuficiencia cardíaca o renal y anemia, que pueden limitar la aplicabilidad de este tratamiento (Spina et al., 2012).

1.4.6. Oclusión arterial retiniana

La oclusión arterial retiniana (OAR) se trata de la obstrucción del flujo sanguíneo de la retina que puede deberse a diferentes causas tales como un émbolo que provoca la oclusión o la formación de un trombo, una vasculitis que provoca la inflamación de la vasculatura retiniana, un daño traumático de la pared del vaso o por un espasmo. La falta de suministro de oxígeno a la retina durante la obstrucción puede provocar una grave pérdida de visión en la zona de la retina isquémica (American Academy of Ophthalmology, 2022). Al igual que la OVR, esta enfermedad se puede clasificar en diferentes tipos dependiendo de la zona de drenaje arterial de la retina que esté afectada, que son los siguientes: oclusión de la arteria central retiniana (OACR), oclusión de la arteria hemirretiniana (OAHR) y oclusión de rama de la arteria retiniana (ORAO). La primera de ellas consiste, tal y como indica su nombre, en la obstrucción repentina de la arteria central de la retina, lo que provoca hipoperfusión retiniana y daño celular progresivo (Farris & Waymack, 2021). La segunda (OAHR) es una variante de la primera que afecta únicamente al hemisferio superior o al inferior de la retina (Priyanka Agnihotri, Smriti Gupta, 2018). Por último, la OACR se produce por la obstrucción de una de las ramas de la arteria central de la retina cuya causa más común se debe a la aparición de émbolos secundarios a placas carótidas o cardíacas (Baumal, 2018).

Entre los síntomas de la enfermedad se encuentra el síndrome isquémico ocular, la inyección conjuntival y neovascularización del segmento anterior y posterior con inflamación en la cámara anterior o vitritis. También es común observar HEs en el fondo del ojo, al igual que con la OVR, o incluso edemas en el nervio óptico (American Academy of Ophthalmology, 2022). En cuanto a los tratamientos frente a la OAR, es recomendable el cribado y tratamiento de los factores de riesgo vasculares en los pacientes, para así lograr evitar el avance de dicha enfermedad. No obstante, la literatura actual sugiere que el tratamiento con activador tisular del plasminógeno intravenoso, que se trata de una proteína proteolítica empleada para la disolución de coágulos de sangre, puede ser eficaz (Mac Grory et al., 2021).

1.4.7. Desprendimiento de retina

El desprendimiento de retina (DR) constituye una afección ocular grave que puede provocar una pérdida de visión permanente. Dicha enfermedad se produce cuando la capa neurosensorial de la retina se desprende de la parte posterior del ojo causando la pérdida de suministro de nutrientes y oxígeno y la consecuente muerte del tejido. Se pueden diferenciar tres categorías dentro del DR: desprendimiento regmatógeno, traccional y exudativo. El DR regmatógeno es el más frecuente y la causa que lo produce es el paso de líquido desde la cavidad vítrea a través de una rotura o desgarro de la retina, hacia el espacio entre la retina sensorial y el epitelio pigmentario retiniano (RPE). El DR traccional es aquel que se produce por la contracción de las membranas proliferativas, lo que provoca la elevación de la retina. Por último, el DR exudativo es el resultado de la acumulación de líquido bajo la retina sensorial causada por otras enfermedades de la retina o la coroides (Blair K & Czyz C. N., 2020). Esta última se trata de una capa de tejido vascular y conectivo del ojo localizada entre la esclerótica (membrana que constituye la capa exterior del globo ocular) y la retina (Lambert-Cheatham et al., 2022; Willoughby et al., 2010).

El rápido diagnóstico y tratamiento son esenciales para evitar la gran morbilidad asociada a esta enfermedad. Los factores que predisponen al DR traccional y DR exudativo son más evidentes y fáciles de diagnosticar a diferencia de los del DR regmatógeno. Algunos de ellos son los siguientes: enfermedades oculares locales tales como la miopía o la retinosquisis, adherencias vítreo-retinianas asociadas al desprendimiento de vítreo posterior, la cirugía de cataratas, retinopatías que afectan a la membrana conocida como hialoide y los traumatismos oculares (Ghazi & Green, 2002). Respecto a los posibles tratamientos, éstos dependen de la gravedad y del tipo de desprendimiento y en base a ello se puede recomendar la cirugía láser, el tratamiento de congelación (criopexia) u otros tipos de cirugía que permitan reparar los desgarros o roturas de la retina. Además, en ciertas situaciones, el oftalmólogo correspondiente podrá recomendar más de uno de estos tratamientos al mismo tiempo (National Eye Institute, 2022b).

1.5. El ojo humano

El ojo es considerado uno de los órganos más complejos del cuerpo humano. En él, se pueden distinguir fundamentalmente tres capas: la región exterior formada por la córnea y la esclerótica, la capa intermedia compuesta por el iris, el cuerpo ciliar y la coroides y, finalmente, la capa interna formada por la retina (Figura 1.3). Las capas oculares están rodeadas a su vez por tres estructuras transparentes conocidas como humor acuoso, vítreo y cristalino. La córnea es la encargada de refractar y

transmitir la luz al cristalino y a la retina y de proteger al ojo frente a infecciones y daños estructurales en las partes más profundas. La esclerótica consiste en una capa de tejido conectivo que mantiene la forma del ojo y le protege de las fuerzas internas y externas. El iris controla la cantidad de luz que llega a la retina y, por tanto, el tamaño de la pupila. El cuerpo ciliar es el lugar donde se produce el agua y es el encargado de controlar la potencia y la forma del cristalino. La coroides es la capa vascular que proporciona los nutrientes y el oxígeno a las capas externas de la retina (Willoughby et al., 2010). Finalmente, respecto a la retina, se detallará con mayor profundidad a continuación puesto que es la estructura ocular en la que se centra el trabajo realizado.

La retina es una de las estructuras fundamentales del ojo humano puesto que de ella depende nuestra visión. Concretamente, se trata de un tejido estratificado que recubre el interior del ojo cuya función es convertir la luz entrante en una señal neuronal adecuada, para que pueda ser posteriormente procesada en la corteza visual del cerebro, de ahí que se considere una extensión del mismo (Abramoff et al., 2010). La anatomía de la retina se puede analizar diferenciando dos niveles, macroscópico y microscópico. En la Figura 1.4, se muestran las principales estructuras oculares a nivel macroscópico, que son las que se detallan a continuación (Abramoff et al., 2010; Snell, 2007).

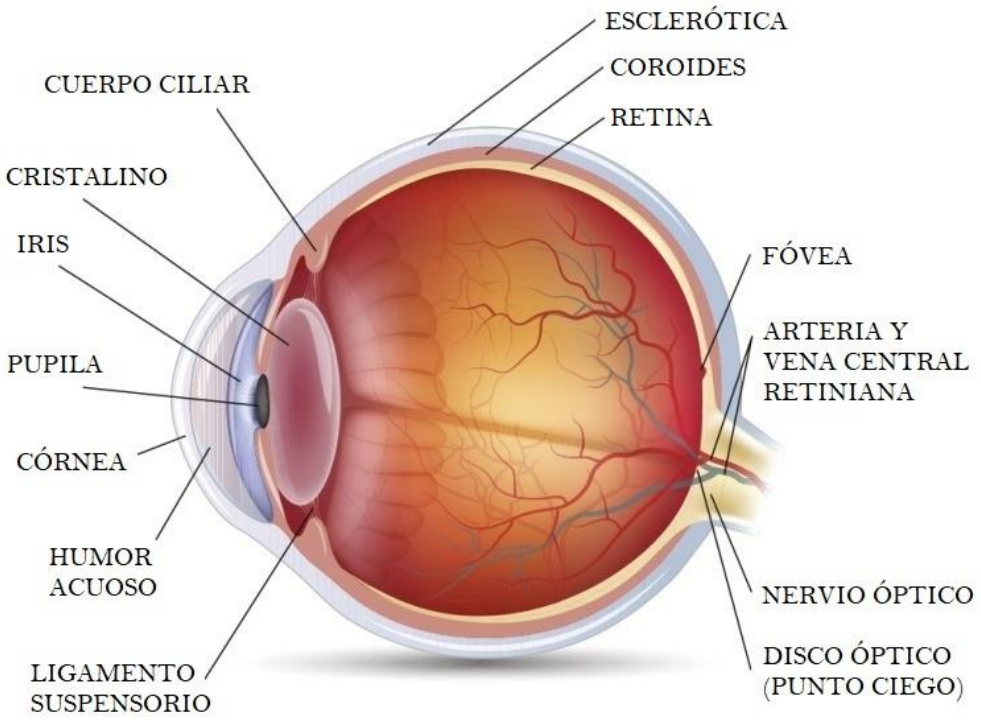


Figura 1.3. Sección transversal del ojo humano donde se muestran sus estructuras principales (Fuente: adaptada de Bixler, 2019).

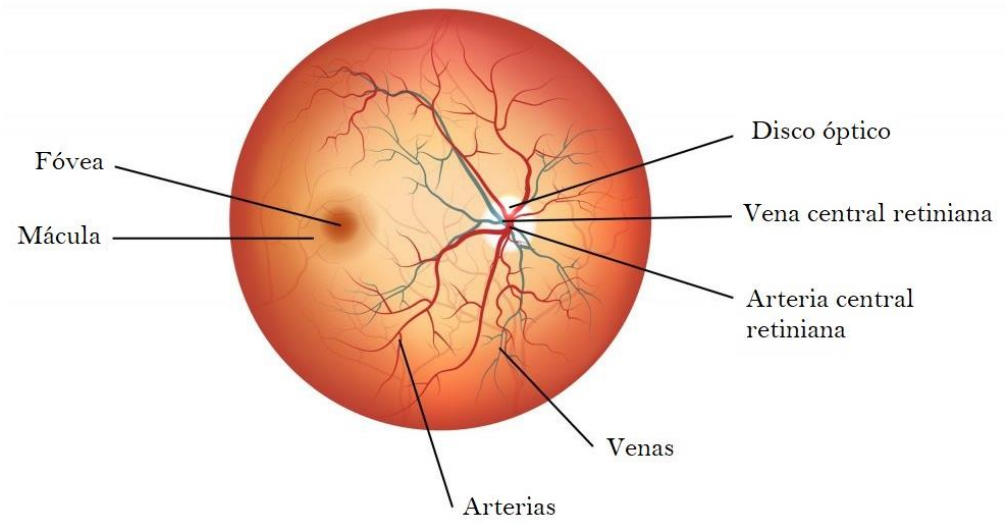


Figura 1.4. Imagen de la retina donde se señalan el disco óptico, la mácula, la fóvea y las arterias y venas (Fuente: adaptada de (Retina - Anatomía Del Ojo - Visión. Definiciones y Conceptos, 2018).

- **Disco óptico (DO).** También conocido como papila o punto ciego (puesto que carece de la presencia de fotorreceptores), es el punto de entrada del nervio óptico en el globo ocular. Dicho de otro modo, es la zona desde la cual salen los axones de las células ganglionares de la retina que forman el nervio óptico. Concretamente, se trata de un disco circular o ligeramente ovalado de 1.5 mm de diámetro aproximadamente y cuyo color es amarillento.

- **Mácula.** Es un área ovalada de alrededor de 5mm de diámetro situada en el centro de la retina que destaca por su gran pigmentación. Es la encargada de la visión central, visión en movimiento y en detalle. Esto último se debe a que es la zona que concentra mayor densidad de células fotorreceptoras, de ahí que nos permita ver con gran agudeza visual, siendo capaces de diferenciar colores y objetos pequeños.

- **Fóvea.** Es una hendidura poco profunda con un diámetro aproximado de 1 mm situada en el centro de la mácula. Contiene únicamente células denominadas conos, que son los fotorreceptores encargados de la percepción de los colores. Se trata del foco receptor de los rayos de luz que llegan a la retina y, por tanto, es considerada el área de mayor agudeza visual.

- **Vasos sanguíneos.** Se consideran vasos sanguíneos de la retina tanto a la arteria como a la vena central y sus ramas. Las arterias y venas se encargan de proporcionar los nutrientes y el oxígeno a la retina. Ambas, la arteria y vena central entran al globo ocular mediante el nervio óptico.

A nivel microscópico, en la retina se pueden distinguir hasta diez capas distintas de neuronas interconectadas por sinapsis, que son las siguientes de menos a más superficiales (Mahabadi Navid & Al Khalili, 2021).

- **Membrana limitante interna (ILM).** La ILM es la superficie interna de la retina que limita con el humor vítreo y que, por tanto, forma una barrera de difusión entre éste y la retina neural. Contiene botones sinápticos de células de Müller (o células gliales) conectados lateralmente.

- **Capa de fibras nerviosas (NFL).** Segunda capa más interna desde el vítreo. En ella podemos encontrar los axones de las células ganglionares que forman el nervio óptico. Estas células son la principal neurona de salida en la retina y entre sus funciones principales se encuentran la liberación de melatonina y la regulación del tamaño de la pupila.

- **Capa de células ganglionares (GCL).** Capa que, como su propio nombre indica, contiene los núcleos de las células ganglionares y las células amacrinas de la retina. En general, las dendritas de las células ganglionares más pequeñas se arborizan en la IPL, mientras que las dendritas de aquellas más grandes se arborizan en otras capas.

- **Capa plexiforme interna (IPL).** Zona compuesta por un denso retículo de fibrillas formado por las dendritas entrelazadas de las células ganglionares y las células de la INL.

- **Capa nuclear interna (INL).** La INL contiene los cuerpos celulares (núcleos) de los cuatro tipos de células, que son las células bipolares, horizontales, de Müller y amacrinas.

- **Capa plexiforme externa (OPL).** Capa que contiene la sinapsis neuronal entre las células fotorreceptoras (conocidas como bastones y conos) y las células horizontales. Éstas últimas intervienen en la modulación de la transferencia de información entre las células bipolares y los fotorreceptores. También, contribuyen a que los ojos sean capaces de adaptarse tanto a condiciones de poca luz como a la luz brillante.

- **Capa nuclear externa (ONL).** Esta capa contiene los gránulos de los conos y bastones que perciben los fotones, los cuerpos celulares de los conos y las prolongaciones de los bastones. En general, los primeros (conos) presentan

núcleos y cuerpos de mayor tamaño que los segundos (bastones).

- **Membrana limitante externa (ELM).** Capa formada por las bases de los cuerpos celulares de las células fotorreceptoras. La ELM forma una barrera entre el espacio subrretiniano (en el que se proyectan los segmentos externos e internos de los conos y bastones para estar en contacto estrecho con el RPE) y la retina neural propiamente dicha.

- **Capa de células fotorreceptoras.** Esta capa está compuesta por los segmentos externos de los fotorreceptores, que son las células de la retina encargadas de transformar la luz en un impulso nervioso. Como ya se ha comentado anteriormente, existen dos tipos denominados conos y bastones. Los conos se concentran en la zona de la mácula y ayudan al cerebro a procesar la visión fotópica, que implica la visión de los colores con diferentes niveles de luz. Por tanto, este tipo de célula es la que nos permite tener gran agudeza visual. Sin embargo, los bastones se localizan en la parte externa de la retina y su densidad aumenta a medida que se avanza hacia la periferia de la misma. Son los encargados de la visión nocturna puesto que su velocidad de respuesta es lenta y su sensibilidad al contraste y agudeza espacial son también muy bajas.

- **Epitelio pigmentado (RPE).** Capa más externa que se compone de células cúbicas que contienen gránulos de melanina. Las principales funciones del RPE son el metabolismo de la vitamina A, el mantenimiento de la barrera sangre-retina y la fagocitosis de los segmentos externos de los fotorreceptores. También, se encarga de la producción de la matriz de mucopolisacáridos que rodea los segmentos externos de la retina y del transporte activo de materiales desde y hacia el RPE.

1.6. Imágenes médicas. Retinografías

La ciencia médica ha experimentado un rápido progreso y la invención de ciertos medicamentos ha beneficiado a la humanidad y a toda la civilización. Sin embargo, antes del tratamiento, la principal necesidad es el diagnóstico adecuado y correcto de las enfermedades y es aquí donde las imágenes médicas desempeñan un papel fundamental (Ganguly et al., 2010).

La imagen médica se refiere a los procesos y técnicas utilizados para crear imágenes del cuerpo humano (o partes del mismo) con diversos fines clínicos. Dado que la calidad de las imágenes médicas afecta al diagnóstico, el procesamiento de éstas se

ha convertido en un punto de vital importancia cuyo objetivo es mejorar la interpretabilidad de los contenidos representados. Esto puede implicar una mejora de la propia imagen para aumentar la percepción de determinadas características, así como la extracción manual o automatizada de información. Dentro del procesamiento de imágenes, existen diferentes técnicas de análisis y visualización que se engloban en las siguientes categorías (Ritter et al., 2011).

- **Realce** (*Image enhancement*). Consiste en la mejora de los contornos de la imagen y otras propiedades relevantes, así como la eliminación de distorsiones tales como el ruido y las inhomogeneidades de fondo.

- **Segmentación** (*Image segmentation*). Es la identificación de los contornos de una estructura anatómica, como por ejemplo un órgano, un vaso o una lesión tumoral.

- **Registro** (*Image registration*). Se trata de la transformación espacial de una imagen para que coincida directamente con la imagen de referencia dada. Esto es completamente necesario, por ejemplo, en la visualización combinada de imágenes de diferentes modalidades.

- **Cuantificación** (*Quantification*). Consiste en la determinación de las propiedades geométricas (tales como, volumen, curvatura y diámetro) y propiedades fisiológicas (composición de los tejidos o características de perfusión) de una estructura anatómica.

- **Visualización** (*Visualization*). Se trata de la renderización bidimensional (2D) y tridimensional (3D) de los datos de imagen y de modelos virtuales de órganos y estructuras anatómicas.

- **Detección asistida por ordenador** (*Computer-aided detection*). Detección y caracterización de estructuras y lesiones patológicas, así como lesiones tumorales u obstrucciones de vasos, entre otros.

Una imagen de fondo de ojo (o imagen retiniana) se define como el proceso mediante el cual se obtiene una representación bidimensional de los tejidos semitransparentes de la retina tridimensional, proyectados en el plano de la imagen mediante luz reflejada (Abramoff et al., 2010). Mediante el uso de este tipo de imágenes, los oftalmólogos podrían detectar a tiempo lesiones o patologías que afectan a la capacidad de visión de un paciente, evitando que acabe en la pérdida

visual completa o ceguera. Su obtención, siempre ha constituido un reto tanto para la óptica como para la física. No obstante, a lo largo de los años, se han ido desarrollando diferentes técnicas o sistemas que permitían obtener dichas imágenes. Entre las técnicas más relevantes se encuentran las que se detallan a continuación en orden cronológico (Cole et al., 2016; Fernández Revuelta, 2012; García Gadañón, 2008; Keane & Sadda, 2014; Riordan-Eva, 2012).

- **Oftalmoscopia.** Es la técnica más utilizada en la práctica clínica que permite observar el fondo del ojo, pero sin obtener una imagen permanente del mismo. Se pueden distinguir dos tipos de oftalmoscopios: directo e indirecto. El oftalmoscopio directo se trata de un instrumento óptico portátil que proporciona una visión del fondo ocular mediante su iluminación con luz proyectada a través de un prisma. Concretamente, la luz se refleja en la retina y es recogida por el observador a través de un orificio situado encima del prisma. El color, el tamaño del área iluminada y el enfoque del oftalmoscopio, se pueden ajustar. Este tipo de instrumento destaca por su sencillez de uso, su portabilidad y su bajo coste, de ahí que se utilice en el examen médico general estándar. No obstante, aunque es cierto que permite obtener una visión detallada del DO y de la vasculatura retiniana, no es posible observar con claridad la periferia de la retina y la percepción de profundidad. Por otro lado, el oftalmoscopio indirecto (también denominado binocular) consiste en un instrumento que se coloca en la cabeza del observador y que envía una gran fuente de luz que se ajusta de tal manera que coincida con el eje de la mirada. El observador puede ver la retina con diferentes aumentos ayudándose de lentes convexas que poseen diferentes dioptrías. También es una técnica sencilla pero, a diferencia de la anterior, la oftalmoscopia indirecta posibilita el examen de toda la retina, más allá de su extremo periférico. Además, el examinador puede emplear sus dos ojos, lo que le permite tener una vista estereoscópica y, por lo tanto, una mayor capacidad para distinguir las posibles lesiones.

- **Angiografía fluoresceínica.** Esta técnica, empleada desde finales de los años 60, requiere la inyección de un contraste denominado fluoresceína sódica en una vena del antebrazo del paciente. La fluoresceína es una sustancia que permite colorear la sangre que circula por los vasos sanguíneos, ya que sus moléculas emiten luz verde al estimularlas con luz azul. Una vez inyectada, esta sustancia primero llega a los vasos coroideos difundiéndose a través de sus capilares y, después, a los vasos retinianos. Por tanto, esta técnica permite observar las estructuras vasculares y anatómicas del fondo ocular resaltadas, para lo cual se emplean ciertos tipos de oftalmoscopios con filtros especiales. Otra técnica similar es la angiografía con indocianina verde (ICG) cuya diferencia principal es la sustancia de contraste que es inyectada.

- **Oftalmoscopia de láser de barrido** (*Scanning Laser Ophthalmoscope,* SLO). Esta técnica surge en los años 80 y permite obtener imágenes retinianas de alta calidad sin necesidad de dilatar la pupila del paciente. Esto se consigue mediante barridos que realiza, a cierta velocidad, un pequeño punto de láser sobre la región reducida de la retina, para así simular que el área está iluminada homogéneamente. El fotomultiplicador es el encargado de capturar la luz reflejada en cada punto de la retina. Se utiliza en aplicaciones clínicas tales como la realización de densitometrías retinianas, el examen de la hemodinámica retiniana, la oftalmoscopia con distintas longitudes de onda y en la evaluación de la topografía de la papila. Con el tiempo, la SLO ha ido evolucionando permitiendo obtener imágenes no midriáticas (es decir, sin la pupila dilatada) del campo de visión con una apertura de hasta 200°.

- **Tomografía de coherencia óptica** (*Optical Coherence Tomography*, OCT). Esta técnica aparece en los años 90 y permite visualizar cortes histológicos de la retina (tales como imágenes del nervio óptico y de la mácula), sin producir molestias al paciente (técnica no invasiva). A diferencia de la SLO, la OCT permite observar defectos ópticos por debajo de la superficie retiniana, por lo que supuso un gran avance para el análisis en vivo de la retina posterior, la mácula, la papila, así como la coroides y el vítreo. El oftalmoscopio OCT consiste en la integración de la SLO y de la OCT y se trata de un instrumento que produce un registro de imágenes pareadas y simultáneas de la superficie retiniana y de capas más profundas situadas por debajo de ésta.

- **Autofluorescencia.** Se trata de una técnica no invasiva que permite observar el fondo ocular mediante la emisión estimulada de luz desde las moléculas (tales como la lipofuscina) del RPE. La molécula de la lipofuscina se trata de una mezcla de pigmentos autofluorescentes que, debido a la degradación incompleta de los segmentos exteriores de las células fotorreceptoras del ojo, se acumulan en el RPE. Mayor será la intensidad de la autofluorescencia cuanto mayor sea la distribución y la cantidad de dichas moléculas. Por tanto, las imágenes que se obtienen con esta técnica permiten mostrar signos de daño oxidativo que revelan cambios metabólicos a nivel del RPE y que son útiles cuando se complementan con otras técnicas clásicas.

- **Retinografía.** Técnica indolora y no invasiva que consiste en la toma de imágenes a color de la retina a través de una cámara fotográfica especial (retinógrafo) que permite registrar imágenes digitales de alta calidad. Con ella, se pueden observar de manera exacta las principales estructuras oculares de la

retina, es decir la papila, la macula y los vasos sanguíneos. Por tanto, es el único método no invasivo que permite realizar un examen de la circulación sanguínea ya que, como ya se ha comentado anteriormente, otras técnicas como la angiografía fluoresceínica requieren la inyección de un contraste. Las retinografías se pueden obtener mediante dos tipos de retinógrafos: midriático y no midriático. La midriasis se define como la dilatación o aumento del diámetro de la pupila mediante la aplicación de un fármaco anticolinérgico denominado tropicamida. Teniendo en cuenta esto, los retinógrafos midriáticos requieren previamente la provocación de midriasis en el paciente, a diferencia de los no midriáticos. Aunque es cierto que los primeros provocan una sensación molesta al paciente (puesto que la midriasis origina una gran sensibilidad a la luz y visión borrosa), permiten obtener imágenes de la superficie retiniana con mayor campo de visión (FOV) que los no midriáticos.

- **Angiografía por tomografía de coherencia óptica** *(Optical Coherence Tomography Angiography,* OCTA). Técnica de imagen relativamente novedosa que permite la visualización tridimensional de los vasos retinianos y de la coroides, sin necesidad de tener que inyectar un medio de contraste. En primer lugar, se genera una imagen mediante el análisis de las señales de descorrelación que se producen por el movimiento de los eritrocitos (o glóbulos rojos) entre las exploraciones individuales. Posteriormente, se aplica el algoritmo de descorrelación de amplitud de espectro dividido, para conseguir reducir la relación señal a ruido y así poder obtener una imagen más limpia. La OCTA supone un área prometedora en la tecnología de imagen de la retina puesto que tiene una aplicación potencial en el cribado, el seguimiento y el tratamiento de la RD. Además, podría remplazar a la angiografía fluoresceínica en la evaluación de una enfermedad vascular retiniana, puesto que permite detectar lesiones en los plexos capilares retinianos superficiales y profundos.

En la actualidad, la técnica más utilizada de entre todas las anteriores es la retinografía. Su análisis está cobrando cada vez mayor interés puesto que el tipo y la gravedad de ciertas lesiones pueden estar asociados también con otras patologías no oculares, tales como enfermedades vasculares (Besenczi et al., 2016). Además, es el método que mejores resultados arroja para la detección automática de la RD (Lu et al., 2021).

No obstante, la retinografía también se emplea para el diagnóstico y seguimiento en serie de otras enfermedades o afecciones oculares, tales como el glaucoma, la ORVR, la OVCR, la DMAE o el DR, entre otras (Cen et al., 2021).

1.7. *Deep learning* para el diagnóstico en imágenes médicas

El aprendizaje automático (o *Machine Learning*, ML) es considerado un subconjunto de la inteligencia artificial (IA) y describe la capacidad de aprendizaje de los sistemas que, con una mínima intervención humana, son capaces de clasificar en categorías o predecir condiciones inciertas o futuras. Generalmente, los datos que se utilizan en los algoritmos de ML se agrupan en tres conjuntos: entrenamiento, validación y test. A partir de los datos de entrenamiento, los sistemas de ML aprenden la distribución de características de los datos de entrada. Posteriormente, se emplean los datos de validación para optimizar los parámetros del algoritmo. Por último, mediante el uso del conjunto de datos de test, se puede evaluar el rendimiento del algoritmo de ML (Kim et al., 2019).

Como parte del ML, se encuentran las redes neuronales artificiales (ANN). Son algoritmos inspirados en el cerebro humano que constan de capas con nodos conectados. La primera capa recibe los datos de entrada y la última proporciona la predicción de la red. Durante el proceso de entrenamiento, el peso de cada nodo del modelo se configura mediante algoritmos de aprendizaje, siendo el más común el de retropropagación (o *backpropagation*). Este algoritmo consiste en optimizar los pesos de cada nodo aplicando la regla de la cadena y calculando el gradiente con el objetivo de reducir las pérdidas (LeCun et al., 1998). Es mediante la iteración del algoritmo de *backpropagation* como se obtienen los pesos optimizados en el modelo. No obstante, las ANNs tienen ciertas limitaciones. A veces surge el problema de sobreajuste por el cual el modelo se adapta demasiado a los datos de entrenamiento, y es incapaz de generalizar. Además, con este tipo de redes es necesario un paso previo de extracción de características de los datos de entrada. Es por ello por lo que el aprendizaje profundo o DL está cobrando cada vez más fuerza, puesto que evita tener que realizar previamente dicha extracción. En general, las redes neuronales profundas (DNNs) muestran un mejor rendimiento en tareas de predicción como la clasificación y regresión, respecto a las ANNs (Kim et al., 2019).

En consecuencia de lo anterior, los algoritmos de DL y, concretamente, las redes neuronales convolucionales (CNN), se están utilizando cada vez más en el análisis de imágenes médicas (Alyoubi et al., 2020). Las CNNs son un tipo de redes neuronales que han demostrado su eficacia en el reconocimiento y en la clasificación de imágenes puesto que amplían a las redes normales añadiendo operaciones de convolución, no linealidad y submuestreo. Gracias a la convolución, es posible extraer características de las imágenes de entrada. También, se pueden realizar ciertos efectos de procesamiento de imagen, tales como la detección de bordes, la nitidez y el desenfoque, al convolucionar las imágenes de entrada con matrices cuadradas específicas (Wang et al., 2018).

Dentro del campo de la oftalmología, el tratamiento de las enfermedades es más eficaz cuando se detectan en una fase temprana. En los pacientes diabéticos, la revisión periódica de la retina es esencial para poder diagnosticar y tratar a tiempo las patologías oculares asociadas, evitando así el riesgo de ceguera. Los enfoques de DL permiten obtener un gran rendimiento en el cribado de enfermedades del ojo a partir de retinografías (Islam et al., 2020; Ting et al., 2019). Por tanto, es en este contexto donde los sistemas automatizados de detección de patologías cobran especial interés, puesto que permiten ahorrar costes y tiempo y son más eficaces que el diagnostico manual. Éste último requiere más esfuerzo puesto que las imágenes digitales retinianas deben ser examinadas por un oftalmólogo u optometrista con experiencia (Alyoubi et al., 2020).

1.8. Inteligencia Artificial Explicable (XAI)

Tal y como ya se ha comentado previamente, los avances en DL prometen mejorar de manera sustancial el cribado de las enfermedades de la retina y la precisión del diagnóstico. Los sistemas desarrollados con estos métodos han demostrado tener una precisión de nivel experto a la hora de diagnosticar y controlar la progresión de enfermedades oculares tales como la RD, la DMAE, el glaucoma y otras anomalías asociadas a afecciones de la retina. Sin embargo, se desconoce el impacto de estos modelos en el ámbito clínico (Muddamsetty et al., 2021).

Para las tareas sensibles que afectan al bienestar y a la salud de las personas, es crucial limitar la posibilidad de que se produzcan decisiones y acciones inseguras, inadecuadas o no robustas. Por tanto, antes de desplegar un sistema que utilice IA, existe una fuerte necesidad de validar su comportamiento para así establecer garantías de que seguirá funcionando tal y como se espera cuando se despliegue en un entorno del mundo real. Los modelos simples, como las curvas de respuesta o los árboles de decisión poco profundos, son fácilmente interpretables pero su capacidad de predicción es limitada. Las redes neuronales más recientes, basadas en el uso de DL, proporcionan una capacidad de predicción muy superior, pero a cambio de comportase como una “caja negra” en la que el razonamiento subyacente es mucho más difícil de extraer. Con este objetivo, se ha desarrollado XAI, un subcampo de la IA centrado en exponer de forma sistemática e interpretable modelos complejos de IA a los humanos (Samek et al., 2019) .

Dentro de XAI, existen varios atributos o características básicas que normalmente no se tienen en cuenta en los algoritmos de ML o DL existentes. Dichas características son las siguientes (Hussain et al., 2021).

- **Explicabilidad *(explainability)*.** Característica activa de un modelo de aprendizaje a través de la cual se pueden describir claramente los procesos llevados a cabo por el modelo. Por tanto, el objetivo de ésta es aclarar el funcionamiento interno del modelo de aprendizaje. Es importante señalar que las aplicaciones críticas necesitan la explicabilidad no solo por mejorar la interpretación de los modelos, sino también por el factor riesgo cuando, por ejemplo, hay vidas humanas en peligro.

- **Interpretabilidad *(interpretability).*** A diferencia de la anterior, es una característica pasiva de un modelo de aprendizaje que permite a los usuarios entender y dar sentido al modelo.

- **Comprensibilidad *(understandability).*** Se refiere a la característica de un modelo de aprendizaje en la que un usuario es capaz de entender la función del modelo sin necesidad de dar explicaciones sobre los procesos internos que ocurren en el modelo. Es similar al termino conocido como inteligibilidad en el contexto de la IA.

- **Transparencia (*transparency*).** La transparencia está directamente relacionada con la comprensibilidad ya que se considera que un modelo de aprendizaje es transparente si muestra comprensibilidad por sí mismo y sin ninguna interfaz. Dicho de otro modo, cuando un modelo de aprendizaje es intrínsecamente comprensible sin la necesidad de introducir componentes adicionales, el modelo es transparente. Dado que engloba a la explicabilidad y a la interpretabilidad, es el aspecto que más domina en la exhaustividad de un modelo de aprendizaje.

En relación con la explicabilidad, se diferencian dos tipos de enfoques XAI para explicar los resultados de las DNNs en las imágenes médicas. Aquellos que emplean métodos estándar basados en la atribución y los que emplean técnicas novedosas, a menudo de dominio específico o de arquitectura (Singh et al., 2020a). La necesidad de tener que asignar un valor de atribución o relevancia a cada característica de entrada de una red, ha llevado al desarrollo de varios métodos de atribución. Dichos métodos tienen por objetivo determinar la contribución de una característica de entrada a la neurona objetivo, que suele ser en un problema de clasificación, la neurona de salida de la clase correcta. Los mapas de atribución consisten en mapas de calor donde se disponen las atribuciones de todas las características de entrada teniendo en cuenta la forma de la muestra de entrada. Si la característica tiene una contribución positiva para la activación de la neurona objetivo, se suele marcar en color rojo, mientras que, si afecta negativamente a la activación, el color empleado

es el azul (Ivanovs et al., 2021). En el caso de las imágenes, se marcan los rasgos o píxeles que proporcionan pruebas positivas y negativas. En la Figura 1.5 se muestran algunos ejemplos de atribuciones para diferentes imágenes de entrada empleando varios métodos de atribución basados en la retropropagación. Este tipo de método se caracteriza por calcular la atribución de todas las características de entrada mediante un único paso hacia delante y hacia atrás a través de la red. Aunque es cierto que en algunos métodos es necesario repetir estos pasos varias veces, el número de repeticiones es independiente del número de características de entrada y es mucho menor a otros métodos, como son los basados en la perturbación. Además, ofrecen un tiempo de ejecución más rápido, pero a cambio de tener una relación más débil entre el resultado y la variación de la salida (Singh et al., 2020a).

Existen ciertos métodos de atribución basados en retropropagación que solo proporcionan pruebas positivas (como, por ejemplo, el método *DeepTaylor*), los cuales pueden ser útiles para un determinado conjunto de tareas. Los métodos que proporcionan tanto pruebas positivas como negativas tienden a tener ruido de alta frecuencia y como consecuencia de ello, en ciertas ocasiones los resultados parecen espurios. Es por ello por lo que la tarea de evaluación de los métodos de atribución es compleja, ya que es difícil discernir entre los errores del modelo y los resultados parecen espurios. Es por ello por lo que la tarea de evaluación de los métodos de atribución es compleja, ya que es difícil discernir entre los errores del modelo y los

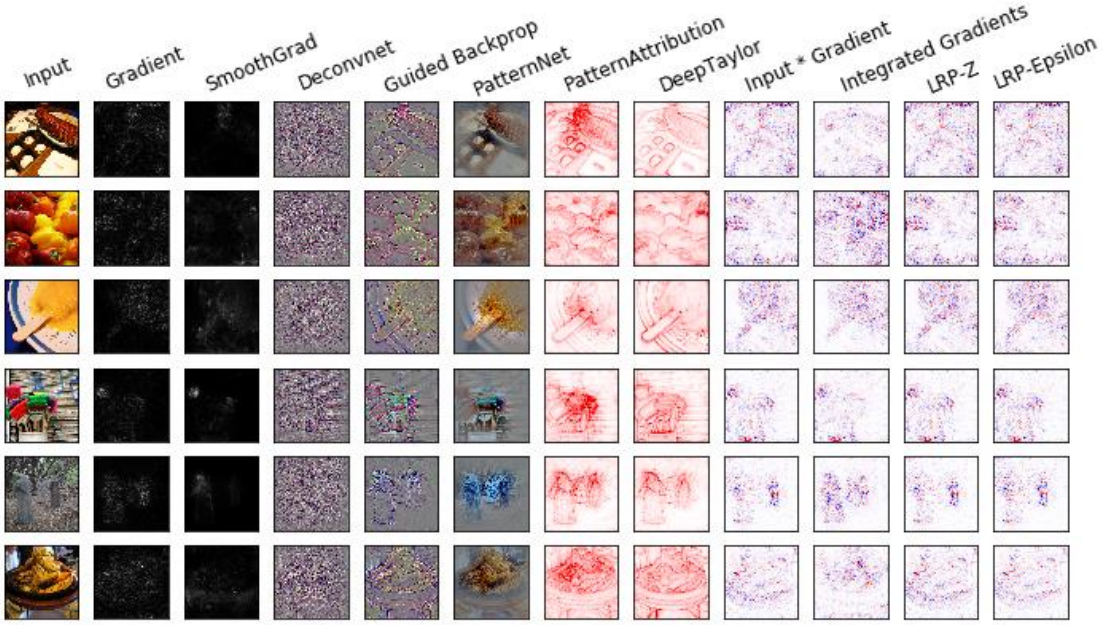


Figura 1.5. Ejemplo de mapas de atribución de una CNN con arquitectura VGG-16 empleando imágenes de Imagenet para diferentes métodos implementados (Fuente: adaptada de Alber et al., 2019).

resultados que ofrece el método de atribución que lo explica. La medida más relevante en la práctica es la similitud de las atribuciones con la expectativa de un observador humano. Esta forma de evaluación de resultados debería realizarse con un experto humano para cada tarea concreta, lo que conlleva un sesgo puesto que los métodos más cercanos a la expectativa del observador pueden verse favorecidos respecto de aquellos que explican el comportamiento del modelo. No obstante, las valoraciones de diferentes métodos de atribución por parte de expertos de un dominio específico, son potencialmente útiles para desarrollar modelos explicables, por lo que debería ser una parte crítica en el desarrollo de un sistema XAI (Singh et al., 2020a).

En el presente trabajo se ha implementado XAI mediante el uso de este tipo de métodos basados en atribución. Esto se debe a que la mayoría de la literatura existente sobre imágenes médicas estudia la interpretabilidad de los modelos de DL empleando dichos métodos (Abeyagunasekera et al., 2022; Singh, et al., 2020b). Su facilidad de uso es una de sus grandes ventajas, ya que se puede aplicar a una CNN sin necesidad de tener que modificar su arquitectura subyacente (Singh et al, 2020a). Concretamente, los métodos que se han implementado son los siguientes: SHAP, *Input x Gradient*, LRP, IG, *SmoothGrad* y *DeepTaylor*, Cada uno de ellos se explicará con mayor detalle en el capítulo correspondiente del documento.

1.9. Hipótesis de trabajo

El envejecimiento de la población, así como el aumento en la esperanza de vida y los cambios desfavorables en el estilo de vida, tales como la disminución del ejercicio físico o los hábitos alimentarios poco saludables, son algunas de las causas del aumento en la prevalencia de las enfermedades que amenazan la visión (Purola et al., 2021). Según la OMS, al menos 2200 millones de personas en todo el mundo padecen problemas de visión de cerca o de lejos y de entre ellas, casi en la mitad de los casos, es decir en 1000 millones de personas, la discapacidad visual se podría haber evitado o aún no se ha tratado. Además, detalla que son siete las principales causas del deterioro de la visión, entre las que se encuentran enfermedades tales como la RD, la DMAE, el glaucoma o las cataratas (World Health Organization, 2021).

Se calcula que la DMAE afecta actualmente a unos 34 millones de personas en la Unión Europea. De entre estos, 22 millones pertenecen a los cinco países europeos más poblados, como son Alemania, Francia, Reino Unido, Italia y España. No obstante, se prevé que el número de pacientes afectados seguirá aumentando hasta llegar a un 25% más en 2050. En el caso de la EOD, la situación actual y futura es

similar. Más del 25% de los pacientes diabéticos padecen alguna patología asociada, lo que supone casi 4 millones de personas (Li et al., 2018). En relación con esta última, la RD está considerada la principal causa de pérdida visual no recuperable en pacientes de entre 20 y 64 años, lo que constituye un 10% de nuevos casos de ceguera cada año. El riesgo de padecer ceguera es 25 veces mayor en pacientes con DM, en comparación con el resto de la población (Aliseda & Berastegui, 2008). Además, estudios realizados han demostrado que la detección y el tratamiento temprano de la RD puede prevenir entre el 50% y el 70% de casos de ceguera (Coney, 2019). No obstante, esto es una tarea difícil puesto que la enfermedad no presenta síntomas distintivos en sus primeros estadios (Oh et al., 2021).

Por las razones anteriores, resulta crucial que los pacientes diabéticos se sometan a revisiones oftalmológicas periódicas que incluyan retinografías (Nentwich & Ulbig, 2015). No obstante, debido a la gran incidencia actual de la DM, esto supondría el análisis de un gran número de imágenes de fondo de ojo, con el consiguiente exceso de trabajo y tiempo para un pequeño número de oftalmólogos especialistas. Por todo lo anterior, los sistemas automáticos de detección y cribado de patologías cobran un gran interés ya que permitirían analizar un número elevado de imágenes.

En el proceso de *screening* de enfermedades oculares, estos sistemas constituirían una etapa previa que evitaría el procesado posterior si no se ha encontrado ningún signo patológico, lo que reduciría considerablemente la carga de trabajo a los expertos. Además, estos sistemas son más eficaces y menos propensos a errores que el diagnostico manual, pues se ha demostrado que los especialistas no siempre coinciden en la valoración del diagnóstico, sobre todo en las etapas intermedias de la enfermedad. Luego, todo ello contribuiría a mejorar la atención sanitaria de los pacientes que presentan algún tipo de patología que afecta a su visión (García Gadañón, 2008; Krause et al., 2017).

En este trabajo se pretende aplicar técnicas XAI sobre un método automático que tiene por objetivo determinar la presencia o ausencia de alguna patología ocular a partir de imágenes de fondo de ojo. De esta manera, sería posible analizar y conocer cuáles son las razones por las que el método ha tomado unas decisiones u otras en su predicción. Para ello, se ha empleado un método basado en una CNN que fue propuesto en (Muñoz Zamarro, 2020), sobre el que se han implementado varios métodos de atribución basados en retropropagación. La diferencia clave respecto a otros estudios ya existentes, es que la CNN empleada se deja de tratar como una "caja negra" y se pretende analizar y justificar sus predicciones.

1.10. Objetivos

Este TFM tiene por objetivo principal aplicar técnicas o métodos XAI para explicar los resultados obtenidos con un método automático de detección de patología ocular basado en una CNN. La detección de patologías en imágenes de fondo de ojo supondría de gran ayuda para los especialistas en el diagnóstico de la RD, ya que disminuiría considerablemente el tiempo y el trabajo en obtener un diagnóstico adecuado. Para conseguir el objetivo principal, se plantearon los siguientes objetivos específicos:

1. Revisión bibliográfica de la literatura existente acerca del análisis y procesado de retinografías empleando métodos de DL junto con técnicas XAI, con el objetivo de tratar de entender los conceptos básicos para la realización del trabajo. Estos conceptos abarcan desde el conocimiento de enfermedades retinianas tales como la RD, hasta la selección de los métodos XAI que se van a emplear en el sistema automático de cribado de patologías. Esta revisión es crítica ya que de ella surgen las bases sobre las que se fundamenta el algoritmo desarrollado.

2. Familiarización con la BD privada de retinografías proporcionada por el Grupo de Ingeniería Biomédica (GIB) de la Universidad de Valladolid. También, se revisaron las bases de datos (BBDD) públicas disponibles con el objetivo de seleccionar aquella que mejor se ajusta al trabajo desarrollado. Dichas BBDD debían incluir un número suficiente de imágenes de pacientes con diferentes patologías oculares. En ambas, las imágenes se han clasificado de manera binaria, es decir en dos categorías: imágenes normales y patológicas.

3. Desarrollo e implementación de diversas técnicas XAI sobre el modelo CNN empleado para la clasificación de las retinografías. Para ello, se ha empleado el lenguaje de programación Python y las librerías, iNNvestigate, SHAP, Keras y Tensorflow. Estas dos últimas librerías de código abierto que permiten adentrarse de manera relativamente sencilla en el mundo del DL.

4. Evaluación de los métodos empleados y obtención de resultados.

5. Verificación del funcionamiento de las técnicas desarrolladas haciendo uso de las BBDD de retinografías seleccionadas.

6. Análisis y discusión de los resultados obtenidos.

7. Extracción de las conclusiones más relevantes del estudio realizado.

1.11. Metodología empleada

En la Figura 1.6 se muestra el diagrama de bloques de la metodologia que se ha seguido para el desarrollo del TFM. En ella se pueden distinguir las siguientes etapas de trabajo.

1. Búsqueda de información y revisión de la literatura con el objetivo de comprender y familiarizarse con el problema que abarca el trabajo. Para lograrlo, se consultaron revistas, libros, artículos científicos y médicos, así como algunas páginas web.

2. Familiarización con el lenguaje de programación Python y aprendizaje de nuevos conceptos relacionados con las técnicas XAI necesarias para implementar el algoritmo correspondiente.

3. Revisión de las BBDD de retinografías empleadas para este TFM.

4. Implementación de los métodos seleccionados para incluir XAI en el método automático de clasificación de retinografías. Para llevar a cabo esta tarea, se emplearon ciertas funcionalidades de Python.

5. Procesado de las retinografías de las diferentes BBDD y obtención de resultados.

6. Análisis de resultados y extracción de las conclusiones más relevantes. Elaboración de la documentación correspondiente del trabajo desarrollado.

1.12. Estructura del documento

El presente documento se organiza en seis capítulos diferentes, cada uno de los cuales se detalla a continuación.

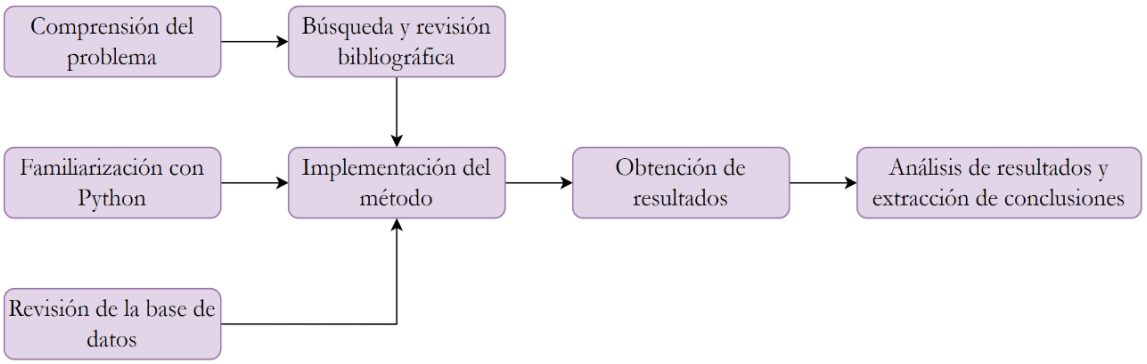


Figura 1.6. Esquema sobre la metodología seguida para la consecución de los objetivos del trabajo.

- **Capítulo 1. Introducción.**

  En este capítulo se han introducido los conceptos básicos en los que se encuentra enmarcado el TFM desarrollado, para así facilitar su comprensión. También, se han mencionado cuáles son las hipótesis del trabajo, así como los objetivos y la metodología seguida para la realización del mismo.

- **Capítulo 2. Revisión del estado de la técnica.**

  En él se realiza una revisión bibliográfica de la literatura existente sobre la tarea de detección automática de patologías oculares a partir de retinografías. Concretamente, se diferencian aquellos métodos basados en técnicas de procesado de imagen, los basados en ANNs y, finalmente, aquellos que emplean CNNs. En estos últimos son en los que se hace un mayor hincapié puesto que son los que más relacionados están con el trabajo realizado.

- **Capítulo 3. Materiales y métodos.**

  En primer lugar, se describen las BBDD que se han empleado y, posteriormente, se explica con detalle el método automático que se ha utilizado junto con los diferentes métodos XAI implementados.

- **Capítulo 4. Resultados**

  En este capítulo se explica el modo de evaluación y se presentan los resultados obtenidos con las BBDD bajo estudio cuando se aplica el algoritmo desarrollado junto con los diferentes métodos XAI.

- **Capítulo 5. Discusión**

  Primero, se interpretan y analizan los resultados obtenidos en este TFM. Luego, se comparan los resultados obtenidos con los diferentes métodos XAI que se han seleccionado y finalmente, se realiza la comparación con los resultados de otros estudios previos.

- **Capítulo 6. Conclusiones y líneas futuras.**

  En el último capítulo se recogen las conclusiones más importantes del trabajo y se detallan las principales aportaciones del estudio realizado, así como las posibles líneas de trabajo sobre las que se podría continuar en el futuro.

Capítulo 2

Revisión del estado de la técnica

2.1. Introducción

En este capítulo se incluye la revisión del estado de la técnica llevado a cabo para la comprensión y el desarrollo de este TFM. Dicha revisión consiste en el análisis y explicación de los diferentes métodos que existen en la literatura y que se utilizan para detectar patologías oculares en imágenes de fondo de ojo. La identificación automatizada de las enfermedades de la retina es un gran paso hacia el diagnóstico precoz y la prevención de la gravedad de la enfermedad. En el pasado se desarrollaron varios métodos que ayudaron a la segmentación e identificación automática de puntos de referencia y patologías de la retina. Sin embargo, los avances actuales sin precedentes en DL y las nuevas modalidades de imagen en oftalmología, han abierto todo un nuevo escenario para los investigadores (Badar et al., 2020). Se comenzará analizando los primeros métodos que se emplearon con este objetivo, que son aquellos basados en técnicas de procesado de imagen. Posteriormente, se pasará a mencionar estudios basados en métodos de ML como son aquellos que emplean ANNs. Por último, se analizarán los métodos más actuales que son los que utilizan técnicas de DL como las CNNs. Respecto a estos últimos, también se mencionarán algunos que integran métodos XAI, que por el momento son muy escasos debido a que son técnicas bastante novedosas. Como ya se ha mencionado anteriormente, tratan de explicar y justificar las decisiones que toma el sistema.

2.2. Métodos basados en técnicas de procesado de imagen

Los recursos informáticos, así como las técnicas de imagen digital, han mejorado a gran velocidad y han ido encontrando cada vez más aplicaciones prácticas. Dentro

del campo de la oftalmología, en los últimos 20 años, los investigadores han empleado técnicas de procesamiento digital de imágenes para el cribado y detección de enfermedades, ya que permiten simplificar el diagnóstico facilitando el trabajo a los oftalmólogos (Zhu et al., 2011). Dichas técnicas de procesamiento de imágenes incluyen la mejora, el registro, la fusión, la segmentación, la extracción de características, la coincidencia de patrones, la clasificación, la morfología, las mediciones estadísticas y el análisis de imágenes (Ravudu et al., 2012).

En cuanto a los métodos existentes en la literatura para detectar patologías o afecciones oculares haciendo uso de este tipo de técnicas, la mayoría de ellos se centran en diagnosticar las enfermedades retinianas más comunes como son la RD, el glaucoma y las cataratas.

Respecto a la RD, ciertos estudios se basan en detectar las lesiones más comunes de la enfermedad, tales como EXs, MAs y HEs. En (Bae et al., 2011) se realizó una investigación sobre el reconocimiento de HEs empleando un esquema híbrido en imágenes de fondo de ojo a color. Para ello se utilizó el algoritmo gaussiano 3D y la técnica conocida como CLAHE para mejorar el contraste de las imágenes. La extracción de candidatos a hemorragias se realizó utilizando un patrón redondeado con correlación cruzada normalizada (NCC). En (Dashtbozorg et al., 2018) se propuso un sistema novedoso y fiable para la detección de MAs utilizando el índice de convergencia local. En la extracción de candidatos se empleó un método de ponderación del gradiente y un enfoque de umbralización iterativo. La clasificación final se realizó mediante un clasificador de muestreo híbrido. También, existen métodos que son capaces de detectar varias lesiones de manera simultánea. Este es el caso del método propuesto en (Seoud et al., 2016) en el que se detectan los dos tipos de lesiones anteriores. Su principal contribución fue la utilización de un nuevo conjunto de características de formas, denominadas *Dynamic Shape Features* (DSF), que no requerían una segmentación precisa de las regiones a clasificar. Los mejores resultados se obtuvieron cuando se clasificaron imágenes sin RD frente a imágenes de todos los estadios de RD. En este contexto, el método propuesto obtuvo un área bajo la curva de características de operación del receptor (ROC), comúnmente conocida como (AUC), de 0.90, una sensibilidad del 93.90% y una especificidad del 50%.

En el caso de la RDP, es posible también encontrar neovascularización. En (Welikala et al., 2014) se llevó a cabo la detección automatizada de nuevos vasos sanguíneos a partir de imágenes de retina. La segmentación de los vasos se realizó aplicando dos enfoques, el operador de línea estándar y un nuevo operador de línea modificado. Finalmente, se realizó una clasificación independiente para cada conjunto de

características utilizando un clasificador de máquina de vectores soporte (SVM).
Para el diagnóstico de patologías de la cabeza del nervio óptico tales como el glaucoma, la segmentación del DO es un paso importante y esencial para crear un marco de referencia. Con respecto a esta enfermedad, existen una variedad de parámetros del ojo muy utilizados para su diagnóstico. Algunos de los más relevantes, son los siguientes: la relación copa-disco (CDR), el borde neural de la retina (NRR), el diámetro de la copa, los vasos sanguíneos de la región inferior, superior, nasal y temporal (ISNT) y la atrofia peripapilar (PPA) (Kumar et al., 2019). En relación con esta última, en (Lu et al., 2012) se propuso PANDORA, un sistema que permitía detectar y cuantificar de forma automatizada las regiones de PPA y del DO en imágenes de fondo de ojo. En el estudio de (Issac et al., 2015) fue necesaria la segmentación del DO y de la copa óptica para la creación de un procedimiento versátil basado en umbrales que mejorase la identificación y agrupación de dicha patología ocular. En este caso, para la clasificación final se utilizó un clasificador SVM con varios *kernels*.

Uno de los métodos más actuales que hacen uso únicamente de técnicas de procesado de imagen para el cribado del glaucoma, es el que se recoge en (Zahoor & Fraz, 2017). En él se propuso la segmentación rápida del DO mediante el uso de la transformada polar. La eliminación del DO de las imágenes de la retina, mediante la transformada circular de Hough (CHT), fue el avance fundamental en la construcción del sistema indicativo para el diagnóstico temprano del glaucoma. Este método permitió obtener una precisión óptima del 99.80%, una sensibilidad del 83.09% y una especificidad del 99.93%

Respecto a las cataratas, en (Kolhe et al., 2007) se propuso un sistema automatizado para la detección de esta enfermedad a partir de imágenes retinianas. En él, se calcularon los coeficientes mediante la transformada *wavelet* discreta (DWT) y la transformada discreta del coseno (DCT). A estos coeficientes posteriormente se les aplicó el análisis de componentes principales (PCA) para llevar a cabo la extracción de características. Un clasificador SVM binario se aplicó a las características extraídas para su posterior clasificación en dos clases, imágenes con cataratas y sin ellas. Mejores resultados se obtuvieron con el método propuesto en (Zheng et al., 2014). El sistema desarrollado para el cribado de las cataratas era muy parecido al del estudio anterior, aunque con ciertas diferencias. Se aplicó la transformada discreta de Fourier bidimensional (DFT) a la imagen retiniana y se empleó el espectro obtenido como característica propia de la patología. Las imágenes se clasificaron en cuatro clases diferentes en función de la severidad (normal, leve, moderada y grave) haciendo uso del algoritmo *AdaBoost*.

También, se pueden encontrar estudios en la literatura que se centran en detectar si la imagen de fondo de ojo se corresponde con haber padecido cataratas anteriormente, puesto que se ha demostrado que puede provocar el surgimiento de otras afecciones oculares. El método propuesto en (Nayak, 2013) emplea técnicas de procesamiento de imágenes para clasificar las retinografías en tres clases diferentes: normales, con cataratas y posteriores a cataratas. Para ello, se creó un sistema parecido a los anteriores, aunque las características que se extrajeron fueron diferentes. Finalmente, se empleó un clasificador SVM para determinar a qué clase pertenecía cada imagen. Los resultados obtenidos fueron clínicamente significativos ya que el método consiguió una precisión media del 88.39%, una sensibilidad del 94% y una especificidad del 93,75%.

En la Tabla 2.1 se muestra un resumen con los datos más importantes de los diferentes métodos basados en técnicas de procesado de imagen.

2.3. Métodos basados en redes neuronales artificiales (ANN)

En el ámbito de la oftalmología clínica, la IA ha demostrado un éxito asombroso ya que permite analizar los datos digitales de forma exhaustiva, no invasiva y rápida. El ML es una rama importante en el campo de la IA, cuyo potencial global para localizar, identificar y clasificar automáticamente las características patológicas de las enfermedades oculares, permite a los oftalmólogos proporcionar un diagnóstico de calidad y facilitar la atención sanitaria personalizada (Tong et al., 2020).

Por las razones anteriores, en la literatura actual existen muchos métodos que emplean ML para el cribado de patologías en retinografías. Además, se suele combinar con técnicas de procesado de imagen. Concretamente, en estos enfoques primero se preprocesan las imágenes utilizando diferentes técnicas de procesado (como, por ejemplo, la extracción del canal verde, la mejora del contraste, el cambio de la dimensionalidad, etc.) y se extraen las características discriminantes. Éstas se emplean para la fase final de clasificación donde ya se determina a qué clase pertenece la imagen de entrada. Las técnicas de ML se emplean en esta última fase, ya que permiten construir el modelo de clasificación, siendo los más comunes el clasificador SVM, k-NN, *AdaBoost*, bosque aleatorio (RF), árbol de decisión (DT) o incluso, clasificadores basados en ANNs. Respecto a estos últimos, se ha demostrado que en la mayoría de los casos son los que mejores resultados arrojan, de ahí su gran utilidad (Ishtiaq et al., 2019).

Patología ocular	Autor/es (año)	Descripción breve del método	Resultados sobre el conjunto de test			
			AUC	Precisión	Sensibilidad	Especificidad
RD	Bae et al. (2011)	Esquema híbrido para detección de HEs	-	-	85.00%	-
	Welikala et al. (2014)	Detección de nuevos vasos sanguíneos con clasificador SVM	0.97	-	100%	90.00%
	Seoud et al. (2016)	Detección de MAs y HEs mediante el uso de DSFs	0.90	-	93.90%	50.00%
	Dashtbozorg et al. (2018)	Detección de MAs empleando un índice de convergencia local	0.90	-	82.00%	-
Glaucoma	Lu et al. (2012)	Detección del DO y regiones de PPA	-	89.47%	83.00%	100%
	Issac et al. (2015)	Segmentación del DO y copa óptica basada en umbrales	-	94.11%	100%	90.00%
	Zahoor & Fraz (2017)	Segmentación rápida del DO mediante la transformada polar	-	99.80%	83.09%	99.93%
Cataratas	Kolhe et al. (2007)	Clasificación de la severidad mediante SVM	-	77.70%	93.00%	77.70%
	Nayak (2013)	Detección de la enfermedad y de una etapa posterior mediante SVM	-	88.39%	94.00%	93.75%
	Zheng et al. (2014)	Clasificación de la severidad mediante *AdaBoost*	-	81.52%	-	-

Tabla 2.1. Comparación de métodos basados en técnicas de procesado de imagen.

Para el diagnóstico de la RD, la presencia de exudados en la retina es el síntoma más característico de la enfermedad. A dichos signos clínicos, se les conoce también como lesiones brillantes y suelen detectarse en los estadios tempranos por lo que son de vital importancia para la tarea de cribado masivo y el seguimiento de la RD (Joshi & Karule, 2018). En (Nayak et al., 2008) se propuso un sistema para la clasificación automática de las imágenes de fondo de ojo en tres clases diferentes: normales, con RDNP y con RDP. Para ello, se aplicaron técnicas de procesamiento morfológico y métodos de análisis de textura con el fin de detectar características, tales como el área de EXs, el área de los vasos sanguíneos y el contraste. El sistema desarrollado en (Hanúsková et al., 2013), también permitía clasificar las imágenes en función de si existían lesiones brillantes en ellas o no. En la etapa final de clasificación se utilizó un tipo concreto de ANN, conocido como perceptrón multicapa (MLP). Con este método se consiguió alcanzar un 93.85% de precisión media.

También, se pueden encontrar estudios capaces de detectar el grado de severidad de la enfermedad. En (Paing et al., 2017), se implementó un sistema que permitía clasificar automáticamente las imágenes de fondo de ojo en cuatro clases: normales, con RD leve, con RD moderada y con RD severa. Para ello, primero se extrajeron de la retina, los vasos sanguíneos, EXs y MAs y posteriormente, se utilizaron como características de entrada a la ANN, el área, el perímetro y el recuento de dichas lesiones. En (Al-Jarrah & Shatnawi, 2017) se propuso un sistema muy parecido al anterior, ya que también permitía clasificar la severidad a partir de la detección de lesiones pero, en este caso, de la RDNP. Para entrenar a la ANN se emplearon dos algoritmos diferentes, el de regularización bayesiana y el de retropropagación. Los mejores resultados se alcanzaron con el primero puesto que se consiguió una precisión del 96.6%.

En cuanto a los estudios que existen en la literatura para detectar el glaucoma mediante el uso de ANNs, algunos de los más relevantes son los siguientes. En (Nayak et al., 2009) se propone un método para la clasificación automática de las imágenes de fondo de ojo en función de si contenían o no dicha patología. Para lograrlo, se calcularon las características típicas tales como la CDR, la distancia entre el centro del DO y la ONH, el diámetro del DO y la relación entre el área de ISNT. En el estudio propuesto en (Soltani et al., 2018) se implementó también un sistema de cribado automático de la enfermedad similar al anterior, aunque con algunas diferencias. Se llevó a cabo el reconocimiento de formas y la identificación del DO y la excavación y se extrajeron los principales parámetros del glaucoma. Respecto a estos últimos, se tuvieron en cuenta tanto los parámetros instrumentales (CDR, regla ISNT y asimetría de los ojos) como los factores de riesgo (edad, sexo, historia genética y origen).

Al igual que con la RD, también existen estudios que se centran en detectar un tipo específico de glaucoma, como es el caso de (Oh et al., 2015). En él se diseñó un método de *screening* que permitía diferenciar los casos de GAA de entre todos los candidatos de glaucoma. Para ello, se crearon cinco modelos de predicción del riesgo de GAA utilizando una regresión logística multivariante y una ANN con diversas variables clínicas. Dichas variables informativas, se seleccionaron mediante un algoritmo de evaluación de subconjuntos de consistencia junto con la validación cruzada para optimizar el rendimiento.

Por último, en cuanto a las cataratas, al igual que con las técnicas de procesado de imagen, existen estudios que tratan de detectar si se ha padecido la enfermedad con anterioridad. Este es el caso de (Acharya et al., 2009), en el que se propone un sistema capaz de clasificar automáticamente las imágenes en tres clases diferentes (normal, con cataratas y posterior a cataratas). Para la detección de características específicas, se aplicó el algoritmo de agrupación *K-means* difuso. Como clasificador se empleó una ANN que se entrenó con el algoritmo de *backpropagation* y que consiguió alcanzar una sensibilidad del 97.7% y una especificidad del 100%. Un artículo reciente relacionado con el cribado de esta patología es el que se recoge en (Shehzad et al., 2020). En él, se propuso un nuevo marco de análisis basado en características de textura para identificar la presencia de cataratas en imágenes de fondo de ojo. Aunque se observó que cada imagen digital contenía 220 características de textura, solo se adquirieron 30 de cada una y se optimizaron mediante la unión de tres enfoques de selección: la probabilidad de error con coeficiente de correlación media, *Fisher* y la información mutua.

También, se pueden encontrar estudios capaces de clasificar la severidad de la enfermedad, como por ejemplo el de (Tawfik et al., 2018). En él se propuso un sistema capaz de clasificar las imágenes de fondo de ojo en función del grado de severidad de cataratas. La extracción de características se realizó mediante la utilización de la transformada *wavelet* combinada con la transformada *log gabor* para formar los vectores de características. Finalmente, en la última fase, los vectores de características se enviaron a una SVM y una ANN para clasificar las imágenes en tres clases: normal, fase inicial de cataratas y fase avanzada.

En la Tabla 2.2 se muestra un resumen de los diferentes métodos basados en el uso de ANNs, que se han comentado previamente. Los métodos se ordenan por patologías y por año.

Patología ocular	Autor/es (año)	Descripción breve del método	Resultados sobre el conjunto de test			
			AUC	Precisión	Sensibilidad	Especificidad
RD	Nayak et al. (2008)	Detección del área de EXs para clasificación del tipo de RD	-	93.00%	90.00%	100%
	Hanúsková et al. (2013)	Detección de lesiones brillantes mediante un MLP	-	93.85%	-	-
	Paing et al. (2017)	Clasificación de la severidad mediante detección de lesiones	-	96.00%	95.00%	-
	Al-Jarrah & Shatnawi (2017)	Clasificación de la severidad de la RDNP mediante detección de HEs, MAs y EXs	-	96.6%	-	-
Glaucoma	Nayak et al. (2009)	Detección automática del DO y de los vasos sanguíneos	-	-	100%	80%
	Oh et al. (2015)	Detección del GAA	0.89	84.00%	78.30%	85.90%
	Soltani et al. (2018)	Detección del DO y de la excavación	-	92.85%	100%	90.00%
Cataratas	Acharya et al., (2009)	Detección de cataratas y *post*-cataratas	-	-	97.70%	100%
	Tawfik et al. (2018)	Detección de la severidad utilizando transformadas	-	92.30%	-	-
	Shehzad et al. (2020)	Detección de patología basada en características de textura	-	99.38%	-	-

Tabla 2.2. Comparación de métodos basados en redes neuronales artificiales.

2.4. Métodos basados en redes neuronales convolucionales (CNN)

Hoy en día, en la mayoría de las tareas de análisis de imágenes médicas se emplean metodologías basadas en DL que hacen uso de CNNs. Las razones de ello se deben a que los enfoques basados en ML tienen algunas limitaciones que el DL consigue suplir. Entre las más relevantes se encuentra la necesidad de intervención por parte de expertos para extraer la región de interés discriminatoria de las imágenes, es decir, sus diferentes características. Por lo tanto, se requiere una enorme cantidad de tiempo y esfuerzos humanos para obtener las características que luego se emplean en la fase final de clasificación (Ishtiaq et al., 2019). Los métodos de DL simplifican el proceso de extracción de características mediante la segmentación automática de las mismas. Es decir, la red se alimenta directamente de las imágenes y se realiza automáticamente el análisis y extracción de características. El procedimiento que se sigue habitualmente es el siguiente. Al igual que con los métodos basados en ML, primero se aplican técnicas de preprocesamiento como, por ejemplo, la corrección de la iluminación, la mejora del contraste y el cambio de tamaño, a las imágenes recogidas con el objetivo de reducir las características ruidosas. Estas imágenes preprocesadas forman las entradas a la arquitectura de DL (que suele ser una CNN) para extraer automáticamente sus características distintivas y a partir de éstas, aprender las reglas de clasificación y sus pesos. Para que este proceso funcione según lo esperado, se requiere una gran cantidad de imágenes con los que entrenar la red. De ahí que, en comparación con los enfoques de ML, el DL requiera una mayor potencia de cálculo y de memoria. No obstante, los estudios que existen en la literatura actual demuestran que los resultados que se obtienen con los métodos basados en CNNs, suelen ser mejores en comparación con el resto, de ahí su gran utilidad (Ishtiaq et al., 2019).

Al igual que los métodos comentados en los apartados anteriores, se pueden distinguir dos principales grupos de algoritmos para la detección de la RD utilizando CNNs: aquellos que tratan de detectar las lesiones características de la RD y aquellos que directamente detectan y clasifican la enfermedad. Un ejemplo del primer tipo sería el estudio de (Wang et al., 2020), en el que se desarrolló un sistema de DL para la detección de lesiones tales como EXs, MAs y HEs. En éste, primero se realizó una evaluación de la calidad de las imágenes, que se dividió en tres categorías para su validación (nitidez, posición y artefactos). Solo las imágenes cuyas puntuaciones de calidad cumplían con ciertos estándares, se incluyeron en el estudio. En (Jiang et al., 2019) se construyó un sistema automático basado en varios modelos de CNNs que era capaz de realizar directamente el cribado de la patología. Concretamente, se utilizaron tres modelos (Inception-V3, ResNet-152 y Inception-ResNet-V2) que se entrenaron de forma independiente. Por último, a cada modelo se le asignó un peso

óptimo y los resultados de cada uno de ellos se fusionaron entre sí mediante el algoritmo de *AdaBoost*.

No obstante, la mayoría de los estudios que se pueden encontrar en la literatura, centran sus esfuerzos en detectar el grado de severidad de la enfermedad. Este es el caso del método que se desarrolló en (Saeed et al., 2021). En él, se implementó un sistema de detección automática basado en el uso de una CNN preentrenada, mediante un método de aprendizaje de transferencia en dos etapas. Concretamente, se empleó un modelo base de CNN (ResNet-152) ya entrenado al que se aplicó la técnica de ajuste fino (*fine-tuning*). Este sistema arrojó muy buenos resultados sin que fuese necesario aplicar previamente ninguna técnica de preprocesamiento a las imágenes de fondo de ojo. Concretamente, los mejores resultados se consiguieron con la base de datos EyePACS, puesto que se alcanzó un AUC igual a 0.98 y una precisión, sensibilidad y especificidad iguales al 99.73%, 96.04% y 99.81%, respectivamente.

Para la evaluación automática del glaucoma a partir de imágenes de fondo de ojo, la mayoría de los algoritmos actuales también hacen uso de CNNs. La razón de ello se debe a que estas redes han demostrado tener una gran capacidad para aprender características altamente discriminativas a partir de las intensidades de los píxeles en bruto. En (Diaz et al., 2019) se emplearon cinco modelos de CNNs diferentes (VGG-16, VGG-19, Inception-V3, Resnet-50 y Xception) entrenados con la base de datos *ImageNet* para la detección automática de la enfermedad. También se aplicó el algoritmo de *fine-tuning* para la tarea de evaluación y clasificación de las imágenes en dos clases: normales y con patología.

También, al igual que con la RD, se pueden encontrar estudios que se centran en la detección de la severidad de esta patología ocular. Un ejemplo de ellos es (Serener & Serte, 2019) donde se implementó un método automático para detectar la enfermedad en una etapa temprana o avanzada. La clasificación se realizó utilizando dos modelos de CNN (ResNet y GoogleNet) que fueron preentrenados con aproximadamente un millón de imágenes del conjunto de datos *ImageNet*. Las imágenes de entrada finalmente se clasificaron en tres clases diferentes: normal (o sano), con glaucoma temprano y con glaucoma avanzado. Los mejores resultados se obtuvieron con la arquitectura GoogleNet que consiguió obtener una precisión en la clasificación del 91%, una sensibilidad del 77%, una especificidad del 98% y un AUC igual a 0.91.

En cuanto a la patología de cataratas, se pueden encontrar muchos estudios en la literatura actual que emplean CNNs para su detección automática a partir de

retinografías. Si no se diagnostica la enfermedad en una fase temprana, puede conducir a la ceguera por lo que, la detección precoz es la mejor manera de controlar el riesgo y evitar una cirugía dolorosa. En (Hossain et al., 2020) se propuso una CNN profunda (DCNN) para clasificar las imágenes de fondo de ojo en dos clases, aquellas sin patología y aquellas patológicas. Una de las peculiaridades de este método es que al emplear una DCNN, no se necesitó una etapa de preprocesamiento de las imágenes ya que su arquitectura profunda permitía capturar características semánticas profundas. Concretamente, la arquitectura elegida para la DCNN fue ResNet-50, que permitió alcanzar muy buenos resultados.

Otros modelos son capaces de detectar la gravedad de la enfermedad. El sistema desarrollado en (Pratap & Kokil, 2019) permitía detectar cuatro estadios diferentes de cataratas (normal, leve, moderado y grave) empleando retinografías. La extracción de características se llevó a cabo mediante el uso de una CNN preentrenada cuya arquitectura elegida fue AlexNet. Por último, las características extraídas se aplicaron a un clasificador SVM que realizó la clasificación final. Este sistema permitió alcanzar una precisión en la clasificación en cuatro etapas igual al 92.91%. En (Simanjuntak et al., 2022) se propuso un sistema muy similar puesto que también se empleó una CNN para realizar la clasificación automática de la patología según su severidad. Concretamente, se consideraron cuatro clases diferentes: normal (o sana), inmadura, madura e hipermadura. Para mejorar el aprendizaje de la red, se empleó el optimizador *Adam*.

En la actualidad, con el aumento de los métodos basados en CNNs, se puede observar un incremento en la demanda de su explicabilidad mediante el uso de técnicas XAI, especialmente en áreas de toma de decisiones de alto riesgo como es el análisis de imágenes médicas (Van der Velden et al., 2022). No obstante, en la literatura actual no existen muchos estudios en los que se aplique directamente algún método XAI a las CNNs, por lo que este campo todavía tiene mucho que ofrecer.

Relacionado con la aplicabilidad de XAI en el diagnóstico de la RD, se ha encontrado el siguiente estudio (Araújo et al., 2020). En él se propuso DR|GRADUATE, un sistema automático basado en DL para la clasificación del grado de severidad de la RD, en el que las decisiones se respaldaban mediante una explicación medicamente interpretable y una estimación de cuánto de cierta era la predicción. Esto último permitía al oftalmólogo medir hasta qué punto debía confiar en dicha decisión. Para ello se utilizó una CNN profunda de otro estudio que ya había demostrado su eficacia en esta tarea específica. Concretamente el sistema se evalúo en términos de 1) el rendimiento de la clasificación de la RD, 2) la estimación de la incertidumbre y 3) la explicabilidad. El primer término se evalúo de manera cuantitativa utilizando como

métrica el coeficiente *kappa* y se analizó el rendimiento por clases. El valor de este coeficiente osciló entre 0.71 y 0.84 en las cinco bases de datos que se utilizaron. Posteriormente, se evalúo la relación entre la incertidumbre y el rendimiento. La estimación de la incertidumbre se comparó en imágenes de buena y mala calidad para así poder comprobar si éstas últimas estaban asociadas a incertidumbres más altas. Finalmente, para la explicabilidad se evaluaron cualitativamente a través de la inspección visual los mapas de calor producido por DR|GRADUATE y cuantitativamente por comparación con las anotaciones de las lesiones realizadas por un oftalmólogo especialista.

En cuanto a la detección del glaucoma, tampoco existen muchos estudios en la literatura actual. No obstante, en (Deperlioglu et al., 2022) se propone una solución híbrida con procesamiento de imágenes y DL que es apoyada con XAI para asegurar una toma de decisiones confiables en el diagnóstico de esta enfermedad. Concretamente para el preprocesamiento, se aplicó tanto ecualización de histogramas como CLAHE con el objetivo de mejorar los datos en las imágenes de fondo de ojo a color. Estos datos fueron utilizados posteriormente por una CNN para realizar el diagnóstico y clasificar las imágenes en función de si eran patológicas o no. La XAI se implementó mediante el mapeo de activación de clases (CAM) que permitía explicaciones basadas en mapas de calor del análisis de imágenes realizado por la CNN. Aunque el rendimiento de la solución se probó con varios conjuntos de datos, los valores medios más altos se obtuvieron con la base de datos ORIGA. En concreto, se alcanzó una precisión del 93.5%, una sensibilidad del 97.7%, una especificidad del 92.6% y un AUC igual a 0.951. La contribución de XAI en este estudio fue mediante un análisis humano, es decir, CAM fue evaluado por algunos profesionales médicos. En concreto, este método mostró una precisión aceptable del 82.73%. Por tanto, este estudio demostró que el uso de XAI para el DL de "caja negra" mejoraba el nivel de confianza de los especialistas.

Finalmente, respecto al cribado automático de las cataratas, no se ha encontrado estudios que integren técnicas XAI para explicar y apoyar las predicciones realizadas por modelos basados en CNNs. Esto demuestra de nuevo lo novedoso que es todavía este paradigma.

En la Tabla 2.3 se muestra un resumen con los datos más importantes de los diferentes métodos basados en el uso de CNNs. Los métodos se ordenan por patologías y por año.

Patología ocular	Autor/es (año)	Descripción breve del método	Resultados sobre el conjunto de test			
			AUC	Precisión	Sensibilidad	Especificidad
RD	Jiang et al. (2019)	Detección automática mediante modelo integrado con AdaBoost	0.95	88.21%	85.57%	90.85%
	Wang et al. (2020)	Detección de MAs, HEs y EXs	-	99.70% (MAs) 98.40% (HEs) 98.10% (EXs)	-	-
	Araújo et al. (2020)	Detección del grado de severidad y aplicación de XAI	-	-	-	-
	Saeed et al. (2021)	Detección del grado de severidad	0.98	99.73%	96.04%	99.81%
Glaucoma	Diaz-Pinto et al. (2019)	Cribado automático mediante detección del DO	0.96	89.77%	85.80%	93.46%
	Serener & Serte (2019)	Detección de la patología en su estadio temprano y avanzado	0.91	91.00%	77.00%	98.00%
	Deperlioglu et al. (2022)	Detección automática y aplicación de XAI	0.95	93.50%	97.70%	92.60%
Cataratas	Pratap & Kokil (2019)	Detección de la severidad mediante CNN junto con SVM	-	92.91%	-	-
	Hossain et al. (2020)	Cribado automático mediante una DCNN	0.99	95.77%	94.43%	98.07%
	Simanjuntak et al. (2022)	Detección del grado de severidad	-	92.00%	-	-

Tabla 2.3. Comparación de métodos basados en redes neuronales convolucionales.

Capítulo 3

Materiales y métodos

3.1. Introducción

Tal y como se ha mostrado en el capítulo anterior, los métodos basados en el uso de CNNs permiten alcanzar muy buenos resultados en la detección automática de patologías oculares. En esta sección, primero se describen las bases de datos de retinografías que se han empleado, así como las técnicas de procesamiento que se han aplicado a las imágenes. Posteriormente, se pasa a explicar con detalle conceptos relacionados con el DL con el objetivo de facilitar la comprensión de las diferentes técnicas y de la red CNN que se han utilizado para este trabajo. Por último, también se mencionan los diferentes métodos XAI que se han aplicado al algoritmo implementado.

3.2. Bases de datos de retinografías

Para la realización de este TFM, se ha hecho uso de dos BBDD de retinografías, una privada y una pública, cuyos detalles se especifican a continuación. Concretamente, se han mezclado todas las imágenes de ambas para conformar la BD que se ha utilizado para entrenar el algoritmo desarrollado. Los datos concretos de cada subconjunto empleado se especifican en la Tabla 3.1.

3.2.1. Base de datos privada

Se trata de una BD proporcionada por miembros oftalmólogos pertenecientes al GIB de la Universidad de Valladolid (UVa). Las imágenes proceden del Programa de Cribado de Retinopatía Diabética del Sistema Público de Salud de Castilla y León (Sacyl), el cual se desarrolló como experiencia piloto en las áreas sanitarias de Palencia y de Valladolid Este y Oeste, gracias a la implantación de equipos de retinografía digital en ocho centros de salud. Este programa no hubiese sido posible

	BD privada		BD pública		TOTAL
	Normales	**Patológicas**	**Normales**	**Patológicas**	
Entrenamiento	400	400	401	1519	**2720**
Validación	50	50	134	506	**740**
Test	50	50	134	506	**740**
TOTAL	**500**	**500**	**669**	**2531**	

Tabla 3.1. Número de imágenes para cada clase y base de datos utilizada.

sin la colaboración del centro de lectura del Instituto Universitario de Oftalmobiología Aplicada (IOBA) de la Universidad de Valladolid y sin la colaboración de los profesionales de distintos niveles asistenciales.

Esta BD está formada por un total de 1000 imágenes que fueron capturadas en formato DICOM (convertido a JPG) con un ángulo de visión de 45º. El formato JPG es de 24 bits (8 bits para cada canal de color) por lo que cada píxel puede tomar 256 valores de intensidad diferentes en el rango de 0 a 255, para cada canal RGB. Además, las imágenes cuentan con dos resoluciones diferentes, es decir, la mayoría de ellas tienen 1956×1934 píxeles, pero también existen algunas con una resolución algo mayor igual a 4288×2848 píxeles. No obstante, todas ellas son de buena calidad y se han dividido de manera pseudoaleatoria (es decir, aleatoriamente dentro de cada categoría) para formar el conjunto de entrenamiento, validación y test. El primero de ellos se emplea para entrenar el modelo, es decir, éste irá aprendiendo las principales características de las imágenes que pertenecen a este grupo, a medida que avanza el proceso. El segundo subconjunto se utiliza para examinar la evolución de los modelos ya entrenados y así poder ajustar sus hiperparámetros. Finalmente, el de test sirve para realizar la evaluación final del modelo ya entrenado y, por tanto, para comprobar si generaliza bien cuando se introducen otras imágenes diferentes.

En nuestro caso, para el tamaño de cada subconjunto, se ha elegido una proporción 80:10:10, es decir, de las 1000 imágenes totales, 800 forman parte del conjunto de entrenamiento y el de validación y test únicamente están formados por 100 imágenes. A su vez, dentro de cada uno de ellos, el 50% del total se corresponden con imágenes sanas (clase 0) y el otro 50% restante son imágenes patológicas (clase

1), tal y como se puede observar en la Tabla 3.1. Por tanto, se ha decidido balancear las poblaciones de cada clase.

3.2.2. Base de datos pública

La otra BD empleada en este trabajo es conocida como RFMiD (*Retinal Fundus Multi-Disease Image Dataset*) (Pachade et al., 2021). Se trata de un nuevo conjunto de datos de imágenes de retina de acceso público creado con el objetivo de impulsar los esfuerzos hacia el desarrollo de un sistema de cribado de retina generalizable (es decir, que sea capaz de detectar, además de las enfermedades oculares frecuentes, patologías raras tales como la oclusión de la arteria central de la retina o la neuropatía óptica isquémica anterior, que a menudo son olvidadas pero que en los últimos años se ha detectado que también amenazan la vista). Esta BD fue organizada conjuntamente con el Simposio Internacional del IEEE sobre imágenes biomédicas (ISBI-2021) en Niza (Francia) y sus datos se encuentran alojados en la plataforma *Biomedical Imaging.* Por el momento, RFMiD es el único conjunto de datos disponible públicamente que constituye una variedad tan amplia de enfermedades que aparecen en entornos clínicos rutinarios (Pachade et al., 2021).

Consiste en 3200 imágenes capturadas con tres cámaras de fondo de ojo diferentes (todas ellas centradas en la mácula o en el DO) con 46 tipos de enfermedades oculares distintas anotadas a través del consenso adjudicado de dos expertos en retina de alto nivel. Las imágenes proceden de miles de exámenes realizados durante el periodo comprendido entre 2009 y 2020. En cuanto a la resolución y el ángulo de visión con el que fueron capturadas, existen 2427 imágenes cuya resolución es 2144×1424 píxeles y su ángulo de visión de 45º, 467 con resolución 4288×2848 y 50º de ángulo de visión y 306 con resolución y ángulo de visión 2048×1536 y 45º, respectivamente. Además, se incluyen imágenes tanto de buena como de mala calidad para que así el sistema sea capaz de clasificar un rango más amplio de datos. El formato empleado en todas ellas es PNG, que también cuenta con una profundidad de 24 bits (Pachade et al., 2021).

En este caso, la proporción elegida ha sido 60:20:20, es decir el 60% del total para el conjunto de entrenamiento y para el de validación y el de test, el 20%. Teniendo en cuenta el número total, esto se traduce en 1920 imágenes de entrenamiento y 640 imágenes tanto para validación como para test. A diferencia de la BD privada, ésta no cuenta con clases balanceadas, es decir en todos los subconjuntos, existen más imágenes de la clase patológica que de la sana, tal y como se observa en la Tabla 3.1. El motivo por el que se ha decidido utilizar esta BD también para entrenar y

validar el modelo, se debe a que la BD privada contiene solo retinografías asociadas a las principales enfermedades oculares. Sin embargo, la BD pública cuenta con más variedad, puesto que incluye también imágenes correspondientes a patologías oculares menos frecuentes. Todo ello con el objetivo de construir un modelo que sea capaz de generalizar correctamente con una gran variedad de tipos de enfermedades o afecciones del ojo (Pachade et al., 2021).

Por tanto, como se ha comentado previamente, se ha decidido mezclar las imágenes de entrenamiento y validación de ambas BBDD con el fin de entrenar un sistema que fuese más robusto, capaz de detectar tanto las patologías más frecuentes como las más raras. Una vez entrenado el modelo, se ha evaluado su rendimiento con los conjuntos de test por separado de cada BBDD. Teniendo esto en cuenta, los conjuntos de datos finales que se han empleado para el modelo de este trabajo son los que se muestran en la Tabla 3.2.

3.3. Preprocesado

Una fase previa clave en el análisis automático de fotografías del fondo del ojo es el preprocesado de las imágenes. Las técnicas de preprocesado están enfocadas a mejorar la diferenciación de las retinografías. La mejora es vital puesto que estas imágenes pueden experimentar los efectos indeseables de una iluminación desigual, un contraste pobre y la presencia de ruido. Todos estos efectos contribuyen a la posibilidad de una detección errónea de las lesiones, lo que se traduce en un mal diagnóstico (Kumar et al., 2019).

Como es habitual en los modelos de DL, las técnicas de preprocesado aplicadas son muy sencillas ya que estos modelos ya funcionan muy bien extrayendo la

	Imágenes normales	**Imágenes patológicas**	**TOTAL**
Entrenamiento	801	1919	**2720**
Validación	184	556	**740**
Test BBDD privada	50	50	**100**
Test BBDD pública	134	506	**640**

Tabla 3.2. Número de imágenes del conjunto de entrenamiento, validación y tests.

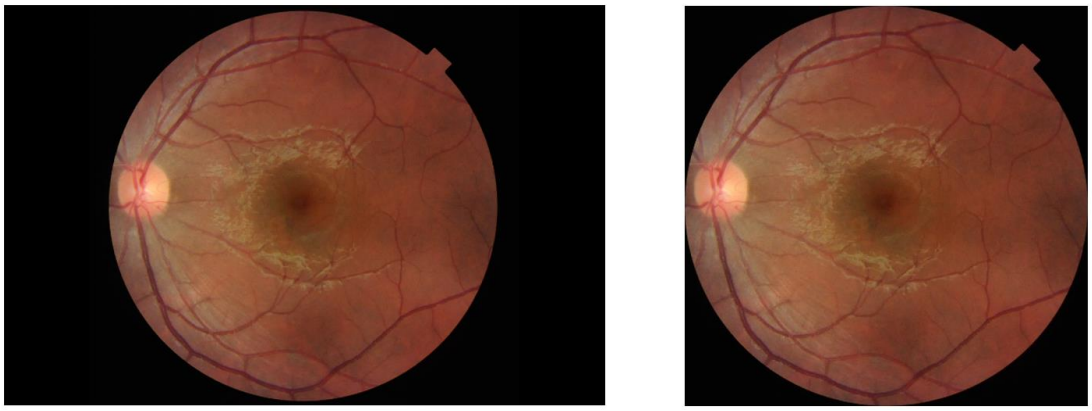
información a partir de las imágenes originales. En este TFM, el preprocesado se ha dividido en tres etapas que son el recorte de la imagen, la normalización del color y la reducción de su resolución, las cuales se detallan a continuación.

3.3.1. Recorte de la imagen

Para eliminar el fondo negro redundante de las retinografías originales, se ha aplicado un método que permite obtener una máscara que delimita la ROI. Con este objetivo, se ha utilizado el canal de luminancia de las imágenes, ya que representa la luminosidad y, para nuestro interés, los píxeles que no son negros. A continuación, se aplicó un filtro de mediana para eliminar el ruido de la imagen (Tobin et al., 2007). Para detectar la ROI, el proceso seguido consistió en iterar todas las columnas y las filas de píxeles desde los bordes de la imagen hasta detectar algún píxel donde la luminancia superase un cierto umbral (píxeles distintos de negro). El nivel de tolerancia para este umbral se estableció empíricamente en 20. La región interior rectangular comprendida entre los píxeles detectados fue nuestra ROI, correspondiente a la FOV. Por lo tanto, se recortaron los bordes de la imagen conservando exclusivamente esta ROI. En la Figura 3.1, se muestra un ejemplo de una imagen original y de una imagen a la que se ha aplicado esta técnica de recorte.

3.3.2. Normalización

Con el objetivo de adaptar la distribución de los datos (en nuestro caso, los valores de los píxeles) a la entrada de la red, las imágenes de las BBDD empleadas se normalizaron en el intervalo $[-1, 1]$ (Romero-Oraá et al., 2019). Al llevar todos los valores a una escala controlada, centrada en cero, se consigue reducir la varianza y, por tanto, se produce una mejora en el proceso de entrenamiento ya que el modelo



(a) (b)

Figura 3.1. (a) Imagen original y (b) imagen recortada.

es más estable (Bishop, 1995). Para lograrlo, dado que los valores de los píxeles están comprendidos entre 0 y 255, fue necesario realizar la siguiente transformación a cada píxel:

$$\left(img * \frac{2}{255}\right) - 1 \tag{3.1}$$

3.3.3. Reducción de la resolución

Dado que los datos de entrada en el sistema desarrollado son imágenes, a mayor número de píxeles, mayor número de operaciones a realizar y, por tanto, mayor número de recursos y tiempo de procesamiento. Por ello, para reducir el coste computacional y, por consiguiente, para acelerar el proceso de entrenamiento, se ha llevado a cabo una reducción de la resolución de todas las imágenes. Concretamente, la resolución empleada en todas ellas ha sido 224×224 píxeles para adaptar el tamaño de las imágenes a la entrada de la arquitectura CNN utilizada en este trabajo (Mishra et al., 2020).

3.4. Redes neuronales

En las últimas décadas, las redes neuronales biológicas han inspirado el desarrollo de multitud de sistemas inteligentes. Muchas disciplinas, como por ejemplo el campo de la medicina, han aprovechado las útiles aplicaciones que pueden ofrecer las ANNs (Haglin et al., 2019). Una ANN es un sistema de procesamiento computacional que está fuertemente inspirado en el funcionamiento de los sistemas nerviosos biológicos, como el cerebro humano. Principalmente, estas redes se componen de un número elevado de nodos, denominados neuronas, que están interconectados entre sí mediante unos pesos y que trabajan de manera distribuida para aprender colectivamente de la señal de entrada con el objetivo de optimizar el resultado final (O'Shea & Nash, 2015).

Una FFNN (o *Feedforward Neural Network*) es la arquitectura más sencilla de una ANN y en ella se pueden distinguir tres tipos de capas: de entrada, ocultas y de salida. Mientras que en todas las ANNs siempre va a existir una única capa de entrada y una de salida, el número de capas ocultas puede variar. Dependiendo de esto último, se dice que se trata de una red superficial o de una red profunda. Aunque es discutible, las arquitecturas con más de dos capas ocultas son habitualmente consideradas como profundas (Emmert-Streib et al., 2020). Un modelo de FFNN muy conocido

que se utiliza principalmente para la clasificación de datos es el MLP. En la Figura 3.2, se muestra un ejemplo de un MLP con una única capa oculta. En ella, $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ representa el vector de entrada que contiene n valores de neuronas de la capa de entrada, $\mathbf{s} = (s_1, s_2, \dots, s_h)^T$ indica el vector de salida formado por h neuronas en la capa oculta y, por último, $\mathbf{o} = (o_1, o_2, \dots, o_m)^T$ representa el vector de salida de m neuronas en la capa de salida. Finalmente, los pesos se muestran mediante $w_{x,y}$, cada uno de ellos hace referencia al peso de la conexión entre el nodo x de la capa anterior y el nodo y de la capa siguiente.

Las ANNs basadas en DL, como son las CNN, son un subtipo de ANNs. Hoy en día, son las CNNs las redes más utilizadas para el reconocimiento de patrones dentro del campo de las imágenes. Su arquitectura está diseñada fundamentalmente para recibir como entradas datos bidimensionales, por lo que permiten codificar características específicas de las imágenes. A su vez, esto hace que la función de transferencia sea más eficiente, por lo que se reduce enormemente el número de parámetros necesarios para configurar el modelo respecto a las ANNs (O'Shea & Nash, 2015). Dado que en el trabajo realizado se emplea una CNN (para luego aplicar los métodos XAI correspondientes), es dentro de este grupo donde se encuadra el algoritmo que se ha implementado.

3.5. Redes neuronales convolucionales

Las CNNs son un tipo especial de FFNN, puesto que en ellas el flujo de información tiene lugar en una sola dirección, desde sus entradas hasta sus salidas. Al igual que

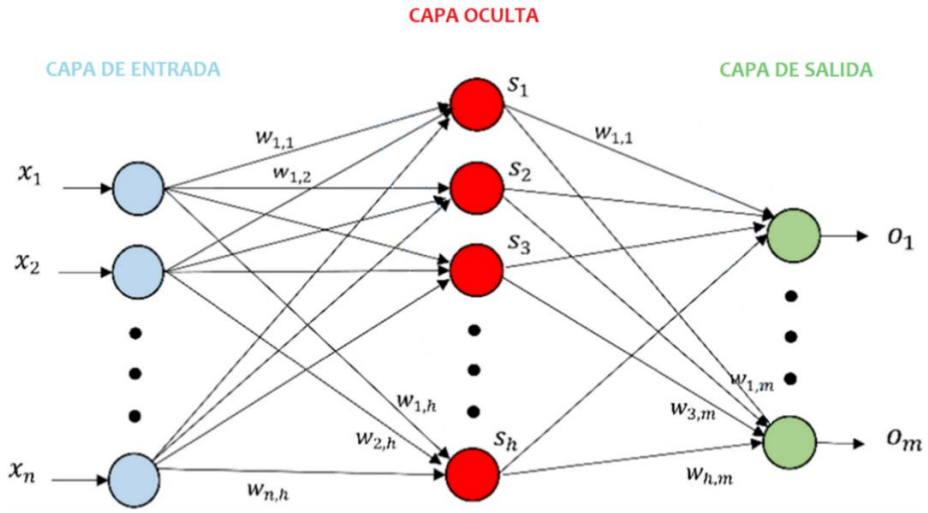


Figura 3.2. Estructura de un MLP con una sola capa oculta (Fuente: adaptada de Pashaei & Pashaei, 2021).

las ANNs convencionales, su estructura se inspira en las neuronas del cerebro humano. Cada neurona de la red lleva asociados unos pesos y un sesgo o factor de corrección, que se emplean para calcular la salida. Concretamente, cuando una neurona recibe varias entradas, se realiza la suma ponderada con el valor de los pesos de cada una de ellas y el resultado es el argumento de una función de activación que genera el valor de salida. Las conexiones locales y los pesos compartidos de una CNN se emplean para aprovechar al máximo las estructuras de datos de entrada 2D, como son las imágenes. No obstante, en los últimos años, gracias a las aportaciones de nuevos investigadores, su uso se ha extendido más allá del procesado de imágenes, demostrando su utilidad en un gran abanico de áreas (Alzubaidi et al., 2021; Rawat & Wang, 2017).

En el entrenamiento de una CNN se pueden diferenciar dos etapas: propagación hacia delante (o *forward propagation*) y propagación hacia atrás (o *backward propagation*). Durante la primera de ellas, se introduce la imagen que se desea clasificar en la red y tras realizar una serie de operaciones, tales como multiplicaciones o convoluciones, se genera una predicción final en la capa de salida. En la CNN empleada en este trabajo, esta predicción se corresponde con la probabilidad de pertenecer a la clase de imágenes sanas o a la clase de imágenes patológicas. Para esta etapa, los parámetros de la red (pesos, sesgos y filtros) se inicializan aleatoriamente y es durante la siguiente etapa, cuando dichos parámetros se van actualizando y optimizando de acuerdo con el error cometido. Concretamente, para calcular este error en la segunda etapa, se emplea una función de coste que permite realizar la comparación entre la salida de la red (que se obtiene con la propagación hacia delante) y el valor esperado. Si el error cometido es elevado, los pesos de la red se actualizan hacia atrás y, por tanto, se consigue que el error vaya disminuyendo a medida que la red se va entrenando. Estas dos etapas anteriores se suceden secuencialmente a lo largo de la fase de entrenamiento hasta lograr un modelo óptimo, cuyos pesos permitan, en nuestro caso, clasificar correctamente las retinografías (Cilimkovic, 2015).

Una de las características esenciales que diferencian a las CNNs respecto a otras redes neuronales es el uso de la operación matemática de la convolución, en vez de la multiplicación matricial. Para realizar esta operación, se aplica un filtro a distintas zonas de la imagen de entrada para generar una salida, que será una mezcla entre dicha entrada y el filtro empleado. Siguiendo esta metodología, el filtro recorrerá todas las posiciones de la matriz de entrada (que se corresponderá con la información que proviene de la capa anterior de la red) para generar una matriz de salida (que dará lugar a la siguiente capa de la CNN). Esto permite utilizar características locales en pequeños grupos de la señal de entrada que acaban proporcionando la información suficiente al modelo para realizar una buena

clasificación. Como consecuencia de ello, en una CNN la conexión entre neuronas es dispersa (*sparse connectivity*), es decir, cada neurona está únicamente conectada a las más cercanas y no a todas las de la capa siguiente, como sí que ocurre en una ANN tradicional. Por tanto, esto supone una gran ventaja puesto que permite que el cómputo de las salidas en una CNN requiera de menos operaciones, dando lugar a una mejora en la eficiencia de los algoritmos (Goodfellow et al., 2016).

Otra ventaja significativa que ofrecen las CNNs frente a las ANNs es la reducción del número de hiperparámetros de la red a entrenar gracias al reparto de pesos (*weight sharing*). En estas redes, las salidas se generan convolucionando un mismo filtro con todas las diferentes zonas de entrada, mientras que, en una ANN, para cada unidad de salida existirán tantos pesos diferentes como número de unidades de entrada. De esta manera, si una capa de una red neuronal tradicional recibe x entradas e y salidas, la red necesitará aprender el valor óptimo de los $x \cdot y$ pesos diferentes para generar las salidas. Sin embargo, en una capa convolucional, la red únicamente tendrá que aprender el valor óptimo de $n \cdot y$ pesos, siendo n el tamaño de los filtros. Luego, como se cumple que n es menor que x, la cantidad de parámetros en una capa convolucional es mucho menor (Goodfellow et al., 2016).

Las últimas ventajas por destacar de las CNNs son la gran cantidad de arquitecturas preentrenadas que existen y que pueden utilizarse para nuevas tareas de reconocimiento (*transfer learning*), la alta precisión en los resultados que se obtienen en la clasificación de imágenes y la equivarianza en las traslaciones de las entradas de una CNN. Esta última propiedad significa que, si se traslada una zona concreta de la señal de entrada a otra localización dentro de la misma entrada, las salidas experimentarán una traslación equivalente (Goodfellow et al., 2016; W. Wang et al., 2019). El motivo por el que se decidió emplear una CNN para desarrollar el algoritmo de este trabajo se debe a todas las ventajas anteriormente comentadas.

3.5.1. Estructura general y capas

En una CNN estándar se pueden distinguir principalmente tres tipos de capas: capa de convolución, capa de agrupamiento (o de *pooling*) y capa completamente conectada (o *fully-connected*). Normalmente, la red está primero compuesta por un número variable de capas convolucionales, cada una de las cuales va seguida por una capa de *pooling* y, por último, se emplean una o dos capas *fully-connected* para generar la salida, que será la probabilidad de pertenecer a cada clase del modelo. Como es lógico, se seleccionará como la clase elegida por el modelo para esa entrada la correspondiente a la salida con mayor probabilidad. A continuación, se va a explicar con mayor detalle cada una de ellas.

3.5.1.1. Capa convolucional

La capa convolucional es el componente esencial de una CNN puesto que es la encargada de realizar la mayor parte del trabajo de cálculo. En pocas palabras, esta capa contiene un conjunto de filtros (o *kernels*) y su función es realizar la operación de convolución entre estos filtros y la señal de entrada a la capa para crear mapas de características (es decir, representaciones de nivel más abstracto). Las características que se recogen en dichos mapas pueden haber sido extraídas por diferentes *kernels*, cada uno de los cuales compartirá a su vez sus pesos con todas las neuronas, tal y como se ha comentado previamente. En una capa convolucional se pueden distinguir diferentes parámetros que intervienen en el proceso de convolución, como son los siguientes (Emmert-Streib et al., 2020; Lopez Pinaya et al., 2019).

- **Número de filtros.** Este parámetro determina el número de detectores de características que se desean tener en el modelo. Es el más variable entre las capas y por lo general, se ajusta a una potencia de 2, entre 32 y 512. Aunque es cierto que el uso de más filtros da lugar a una red más potente, también aumenta el riesgo de sobreajuste debido al mayor número de parámetros que se deben estimar.

- **Tamaño del filtro.** Define la altura y anchura del filtro y, por lo tanto, su extensión espacial, es decir, el tamaño de la región de la señal de entrada con la que se realizará la convolución. Normalmente, se emplean filtros de pequeñas dimensiones, siendo habituales aquellos cuyo tamaño es 3×3, 5×5 o 7×7. La razón por la que, en general, se utilizan filtros pequeños se debe a dos ventajas fundamentales: (1) el número de parámetros que se necesitan aprender se reduce significativamente y (2) se asegura que los patrones distintivos se aprenden de las regiones locales.

- ***Stride*.** El tamaño del *stride* indica el número de píxeles en los que se mueve la ventana del filtro. Su valor habitual suele ser igual a 1, lo que significa que el filtro se deslizará de píxel en píxel. No obstante, se puede utilizar un valor mayor para reducir la dimensión de los mapas de características (Figura 3.3). Esto último se debe a que, cuanto mayor sea este parámetro, menor es el número de convoluciones que se deben calcular y, por consiguiente, menor es el tiempo que tardará el filtro en recorrer toda la matriz de entrada.

- ***Padding***. Parámetro que determina el número de píxeles con valor cero que se desean colocar alrededor del borde de la entrada, de ahí que también se le

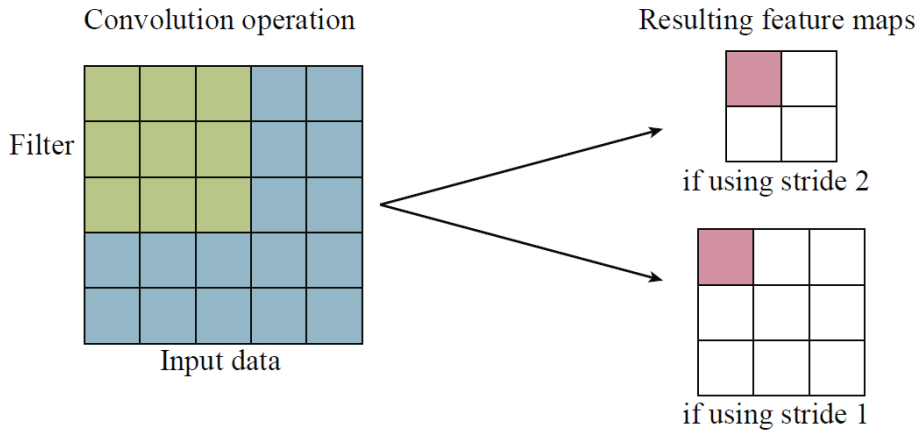


Figura 3.3. Impacto del tamaño del stride en el mapa de características de una capa convolucional (Fuente: Lopez Pinaya et al., 2019).

denomine *zero-padding*. Este efecto evita que el mapa de características se reduzca durante la operación de convolución, ya que el píxel central del filtro puede situarse en el píxel del borde de la imagen de entrada (Figura 3.4). Luego, emplear un *padding* distinto de cero permite conservar la dimensión de los datos y, por tanto, diseñar redes más profundas.

Estos tres últimos hiperparámetros son los que se emplean comúnmente para calcular el volumen de salida de una capa convolucional. Concretamente, para una entrada con dimensiones $W_{in} * H_{in} * Z_{in}$, la salida (es decir, el mapa de características) tendrá dimensiones $W_{out} * H_{out} * Z_{out}$, cuyo valor se calculará dependiendo del tamaño del filtro (N), el *stride* (S) y el *padding* (P).

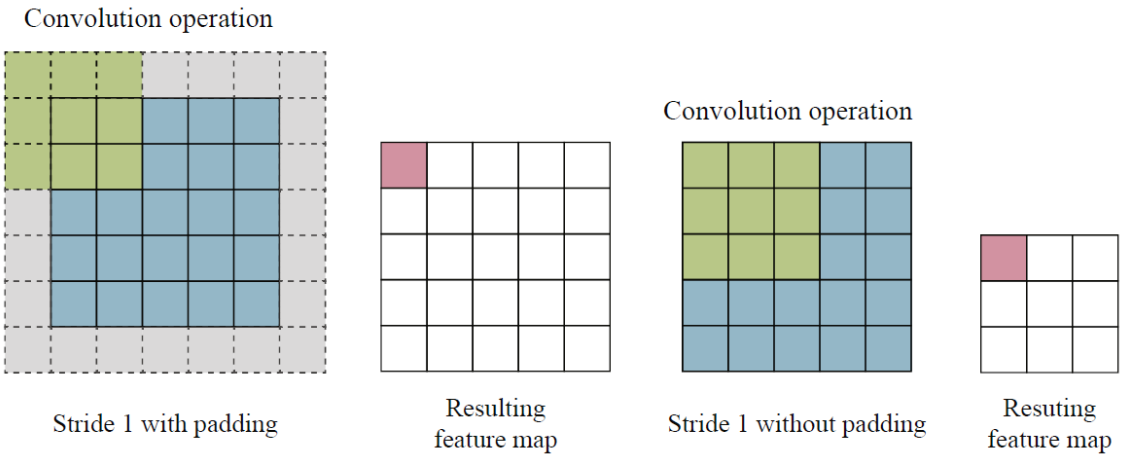


Figura 3.4. Impacto de emplear padding (área gris) alrededor de la entrada de una capa convolucional (Fuente: Lopez Pinaya et al., 2019).

Las expresiones concretas para calcular dichas dimensiones de salida son:

$$W_{out} = \frac{W_{in} - N + 2P}{S + 1} \tag{3.2}$$

$$H_{out} = \frac{H_{in} - N + 2P}{S + 1} \tag{3.3}$$

$$Z_{out} = Z_{in} \tag{3.4}$$

Los valores de los parámetros anteriormente comentados que se han empleado para el algoritmo de este trabajo dependen únicamente de la arquitectura CNN utilizada (se explicará con mayor detalle posteriormente), puesto que no se han añadido capas convolucionales extras. Además, a los resultados que se obtienen tras convolucionar el *kernel* con la entrada, habitualmente se les aplica una función de activación no lineal que permite que la red sea capaz de modelar funciones de gran complejidad puesto que la salida deja de ser una simple combinación lineal. Las funciones de activación más comunes se explicarán en el siguiente apartado.

3.5.1.2. Capa de agrupamiento

Capa cuyo propósito es disminuir el tamaño espacial de la representación capturada por la capa convolucional, de ahí que normalmente se sitúe entre dos capas convolutivas. Principalmente, simplifica la información recogida y crea una versión condensada de la misma información, para lo cual se emplea un método o tipo de *pooling*. Existen diferentes tipos de *pooling*, siendo los más comunes los siguientes (Alzubaidi et al., 2021; Emmert-Streib et al., 2020; Lopez Pinaya et al., 2019).

- ***Max-pooling*.** Es el tipo más común y de forma similar a una capa convolucional, se especifica el tamaño de la ventana (análogo al tamaño del filtro) que se deslizará sobre la entrada para tomar el valor máximo de dicha ventana descartando todos los demás valores.

- ***Average-pooling*.** Se calcula el valor medio de los valores agrupados por la ventana definida.

- ***Global-average-pooling*.** Este tipo está diseñado para sustituir a las capas totalmente conectadas en las CNN clásicas ya que no se define ningún tamaño de ventana, sino que se calcula el valor medio de todos los valores de la entrada.

Es decir, en otras palabras, se toma la media de cada mapa de características generado.

En la Figura 3.5 se muestra un ejemplo concreto para cada tipo de agrupación anteriormente comentado.

3.5.1.3. Capa completamente conectada

De manera general, las últimas capas de una CNN estándar son capas densas o totalmente conectadas. Estas capas son equivalentes a las capas ocultas de una ANN, es decir, conectan cada neurona en una capa con todas las salidas de la capa anterior (Aggarwal, 2018). Su función principal es realizar la clasificación de las características extraídas por la serie de capas convolucionales y capas de agrupación, por lo que está formada por tantos nodos como clases se pretendan clasificar. Para el algoritmo de este TFM, esta capa está formada únicamente por dos neuronas puesto que solo existen dos clases correspondientes a imágenes sanas e imágenes patológicas. Los mapas de características se necesitan aplanar a un único vector 1D antes de ser introducidos en este tipo de capas (Lopez Pinaya et al., 2019).

3.5.2. Funciones de activación

Tal y como ya se ha comentado, a los valores de las operaciones de convolución entre el filtro y la entrada se les aplica una función de activación no lineal. Esto significa que estas funciones toman la decisión de disparar o no una neurona con referencia a una entrada particular, creando la salida correspondiente. Luego, se asemeja al funcionamiento biológico del cerebro donde las neuronas se activan

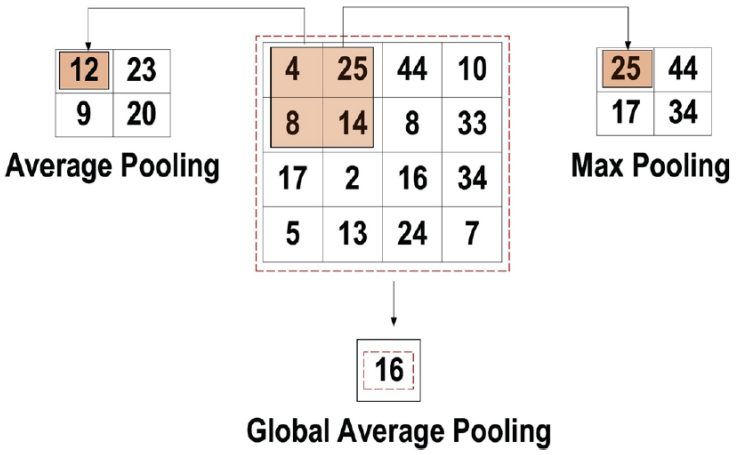


Figura 3.5. Ejemplos de los métodos de pooling más comunes (Fuente: Alzubaidi et al., 2021).

cuando reciben un estímulo que reúne ciertas características. Dentro de este contexto, se pueden diferenciar varios tipos de funciones de activación, siendo las más utilizadas las que se detallan a continuación (Alzubaidi et al., 2021; Emmert-Streib et al., 2020).

- **Sigmoide (*Sigmoid*).** La entrada de esta función de activación son números reales, mientras que la salida está acotada entre 0 y 1 (como pueden ser probabilidades o imágenes normalizadas). Conviene señalar que, si el número de capas de la red es elevado, el gradiente de esta función disminuirá lentamente por lo que el aprendizaje de la red será también muy lento. La curva de la función sigmoide tiene forma de S y puede representarse matemáticamente mediante la ecuación (3.5).

$$f(x)_{sigmoid} = \frac{1}{1 + e^{-x}} \tag{3.5}$$

- ***Softmax*.** Función de activación empleada para calcular la distribución de probabilidad de un vector de números reales. Produce una salida cuyo valor está comprendiendo entre 0 y 1, siendo la suma de todas las probabilidades igual a 1. Esta función se utiliza en modelos multiclase en los que se obtiene la probabilidad de que el dato de entrada pertenezca a cada clase bajo estudio. La clase objetivo será la que obtiene una mayor probabilidad. Por tanto, la principal diferencia con la función anterior es que la sigmoide se emplea en problemas de clasificación binaria, mientras que ésta se utiliza en tareas de clasificación multiclase. Su expresión matemática se especifica en la siguiente ecuación (3).

$$f(x_i)_{softmax} = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \tag{3.6}$$

- **Tangente hiperbólica (Tanh).** Similar a la función sigmoidea puesto que su entrada son también números reales. No obstante, su salida ahora está restringida a valores comprendidos entre el rango -1 y 1. Esta función no lineal puede conducir a gradientes de fuga y es por eso por lo que se emplea en escasas ocasiones en las capas intermedias de la red. Su representación matemática se muestra en la siguiente ecuación (3.7)

$$f(x)_{tanh} = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{3.7}$$

- **Unidad lineal rectificada (ReLU).** Es la función de activación más empleada en el contexto de las CNNs. Convierte los valores enteros de la entrada en números positivos. La principal ventaja que presenta es su baja carga computacional. Además, su no linealidad generalmente conduce a una convergencia más rápida en comparación con las otras dos funciones anteriores. No obstante, ocasionalmente puede darse el problema conocido como "*Dying ReLU*". Este consiste en que se anulan un gran número de neuronas de la red como consecuencia de que todos los valores negativos proporcionan una salida nula. Existen algunas funciones alternativas que solucionan este problema como la que se van a comentar a continuación. La representación matemática de esta función de activación se especifica en la ecuación (3.8).

$$f(x)_{ReLU} = \max(0, x) \tag{3.8}$$

- **Unidad lineal rectificada con fugas (*Leaky* ReLU).** A diferencia de la ReLU, en vez de reducir las entradas negativas, esta función de activación garantiza que esas entradas nunca se ignoren. Por tanto, soluciona el problema del *Dying ReLU* ya que ahora se define un parámetro conocido como factor de fuga (m) para los valores de entrada negativos. Aunque dicho parámetro está acotado entre 0 y 1, comúnmente se establece en un valor muy pequeño, como por ejemplo 0.001. Su expresión matemática se muestra en la siguiente ecuación (3).

$$f(x)_{LeakyReLU} = \begin{cases} x, & si\ x > 0 \\ m * x, & si\ x \leq 0 \end{cases} \tag{3.9}$$

En la Figura 3.6 se muestra la representación gráfica correspondiente a cada una de las funciones de activación mencionadas anteriormente.

3.5.3. Arquitecturas CNN

Desde finales de 1990, la arquitectura y metodología de aprendizaje de las CNN tradicionales han sufrido mejoras con el objetivo de mejorar su rendimiento y hacer las redes más escalables para poder abordar problemas complejos, heterogéneos y de múltiples clases. Entre los principales factores que han contribuido al avance en la investigación de las CNNs, se encuentran la optimización del *hardware* y la disponibilidad de amplios datos de entrenamiento. No obstante, las fuerzas motrices que han permitido las diferentes innovaciones han sido el desarrollo de nuevas estrategias de optimización de hiperparámetros así como el diseño de nuevos

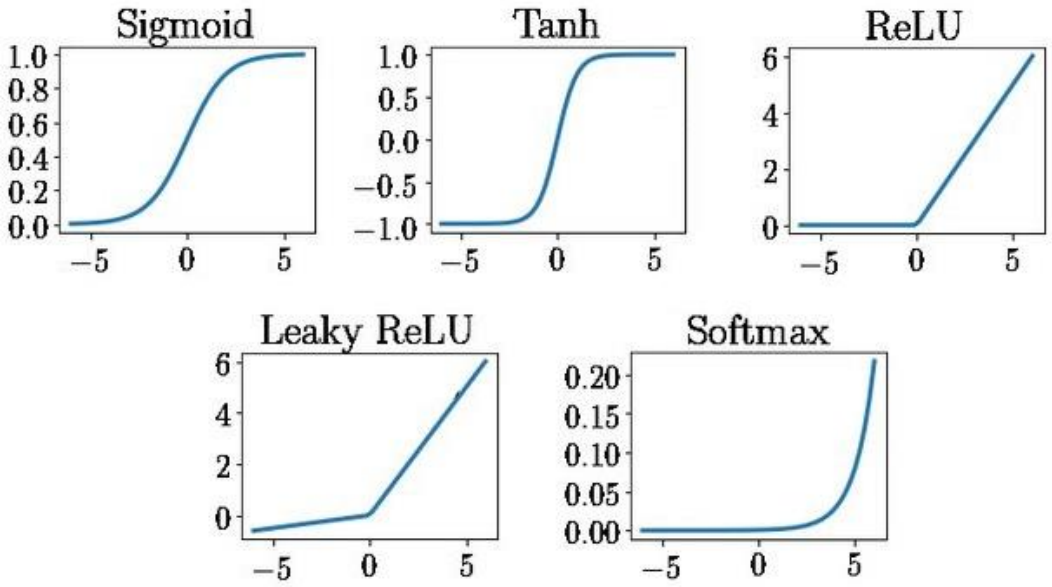


Figura 3.6. Funciones de activación no lineales más comunes (Fuente: Johnson et al., 2020).

patrones de conectividad entre capas y la modificación de las unidades de procesamiento que conforman las redes (Gu et al., 2018; Khan et al., 2020).

En la actualidad, existe una gran cantidad de arquitecturas diferentes, cuya complejidad tanto computacional como arquitectónica (profundidad y número de parámetros a estimar) difiere entre sí, así como los resultados que arrojan. La capacidad de aprendizaje de una CNN suele depender del número de capas o profundidad que tenga la red. El algoritmo más común para entrenar redes neuronales es el conocido como descenso del gradiente (cálculo que permite conocer cómo ajustar los parámetros de la red teniendo en cuenta el error a la salida). A medida que la profundidad aumenta, dicho gradiente irá disminuyendo más lentamente hacia valores muy pequeños y, en ocasiones, esto supone un problema puesto que ralentiza considerablemente el aprendizaje de la red. En el peor de los casos, esto podría llegar a impedir que la CNN complete el proceso de entrenamiento. Diferentes estudios demostraron que existían ciertas arquitecturas, tales como VGG, ResNet o DenseNet, que disminuían este problema del descenso del gradiente y, a su vez, permitían arrojar buenos resultados en las tareas de clasificación y reconocimiento. También, el equipo de Google fue capaz de desarrollar una arquitectura (denominada GoogleNet) que fue la primera en abordar el problema de la gran cantidad de recursos que eran necesarios para entrenar una red neuronal. El consumo de memoria era cada vez mayor, a medida que las CNNs se hacían más profundas. Con esta última arquitectura, se demostró que las capas de la red no siempre tenían que apilarse de manera secuencial, sino que se podía mejorar el rendimiento aumentando la anchura de la red. A partir de

esta idea, surgieron otras arquitecturas que posteriormente se han empleado en muchos estudios de la literatura. Concretamente, algunos de estos modelos son Inception-V2, Inception-V3, Inception-ResNet, entre otros.

En la Figura 3.7, se muestra un diagrama de bolas obtenido del estudio de (Bianco et al., 2018). En él se representa la precisión de distintas arquitecturas CNN (medidas en el conjunto de validación ImageNet-1k para la tarea de clasificación de imágenes) frente a su complejidad computacional medida en operaciones en coma flotante por segundo (FLOPS). El tamaño de cada bola corresponde a la complejidad del modelo medida en millones de parámetros. Tal y como se puede comprobar en la figura, el modelo NASNet-A-Large es el que alcanza la mayor precisión pero, a su vez, es el que tiene la mayor complejidad computacional. Entre los modelos que tienen la menor complejidad computacional (es decir, menos de 5 FLOPS), SE-ResNeXt- 50 (32×4d) es el que proporciona la mayor precisión mostrando al mismo tiempo un bajo nivel de complejidad del modelo (con 2.76 M de parámetros aproximadamente). En general, no se puede inferir una relación clara entre la complejidad computacional y la precisión en la clasificación. Por ejemplo, SENet-154 necesita más o menos el triple de operaciones que SE-ResNeXt-101 (32×4d) y ambos tienen prácticamente la misma precisión. Tampoco existe una relación entre la complejidad del modelo y la precisión puesto que, por ejemplo, VGG-13 tiene un nivel de complejidad (tamaño de la bola) mucho mayor que ResNet-18, mientras que alcanza casi la misma precisión.

En este trabajo, se ha decidido utilizar el modelo CNN desarrollado en nuestro trabajo previo (Muñoz Zamarro, 2020), basado en la arquitectura DenseNet-121. Esta arquitectura arrojó los mejores resultados para nuestro objetivo, en términos de precisión, sensibilidad, especificidad y tiempo medio de evaluación de cada imagen, entre todas las arquitecturas exploradas. En la Figura 3.7, también se puede observar que este modelo es uno de los que ofrece un mejor compromiso entre los tres parámetros analizados. La razón de esto se debe a que es un modelo que ofrece una mayor precisión en la clasificación con una complejidad computacional y del modelo muy pequeña. A continuación, se explica en mayor detalle esta arquitectura.

3.5.3.1. DenseNet

Al igual que ResNet, esta arquitectura fue desarrollada en torno al 2016 con el objetivo de solventar el problema del descenso o desvanecimiento del gradiente. El problema que presentaba ResNet era que preservaba explícitamente la información a través de transformaciones aditivas de identidad, debido a lo cual muchas capas

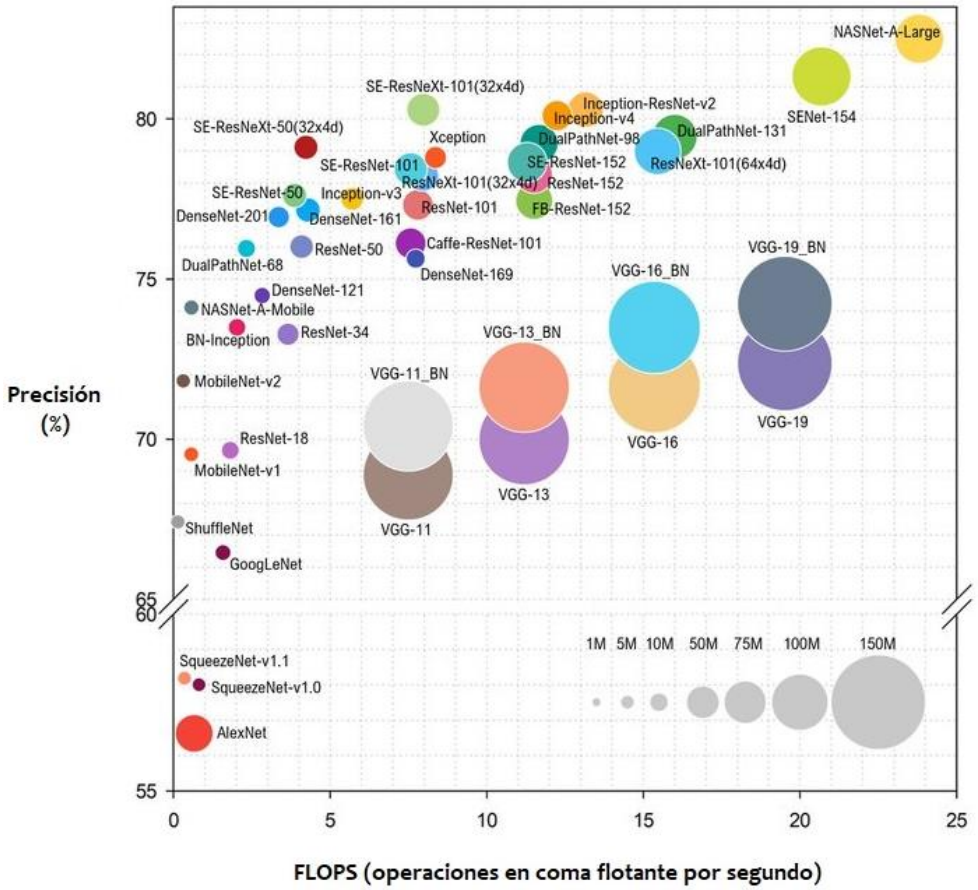


Figura 3.7. Diagrama de bolas que muestra la precisión de cada arquitectura CNN frente a su complejidad computacional medida en FLOPS (Fuente: adaptada de (Bianco et al., 2018).

podían aportar muy poca o ninguna información. Para tratar de abordar esta limitación, DenseNet aportó una modificación de la conectividad entre capas: cada capa precedente se conectaba con la siguiente de una forma *feed-forward*. De esta manera, los mapas de características de todas las capas anteriores se empleaban como entradas en todas las capas posteriores, por lo que las capas recibían un "conocimiento colectivo" de todas sus anteriores. A su vez, esto permitía que el error se pudiera propagar fácilmente y de forma más directa hacía las capas precedentes (Khan et al., 2020).

Todo lo anterior implicaba que, en vez de tener, como en toda CNN tradicional, N capas con N conexiones (una entre cada capa y su siguiente), en DenseNet el número de conexiones directas ahora era igual a $\frac{N\cdot(N+1)}{2}$. En cuanto a su profundidad o número de capas, este tipo de redes se dividían en bloques densos (*dense blocks*), de ahí su nombre. Dentro de cada bloque, las dimensiones de los mapas de características permanecían constantes, mientras que el número de filtros variaba. Además, para unir varios de estos bloques, se empleaban capas de transición que se

encargaban de reducir la dimensión aplicando diferentes operaciones, tales como una normalización por lotes, una convolución 1×1 y una agrupación de capas 2×2 (Khan et al., 2020). En concreto, en este TFM se ha empleado la red DenseNet-121 que cuenta con una profundidad total de 121 capas (Figura 3.8). Este modelo ha sido empleado con éxito en estudios de diversos autores para la tarea de clasificación de diferentes patologías oculares a partir de imágenes de fondo de ojo (Phasuk et al., 2019; Zhang et al., 2021).

3.6. Modelo desarrollado

Tras explicar los conceptos básicos del funcionamiento y de la estructura de las redes neuronales profundas, se pasan a explicar ahora los detalles respecto a la arquitectura CNN empleada, así como las técnicas de DL aplicadas al modelo. El sistema desarrollado se trata de un clasificador automático que permite detectar, a partir de una retinografía, si contiene algún tipo de patología ocular o si se corresponde con un paciente sano. En los sistemas de *screening* de la RD, la

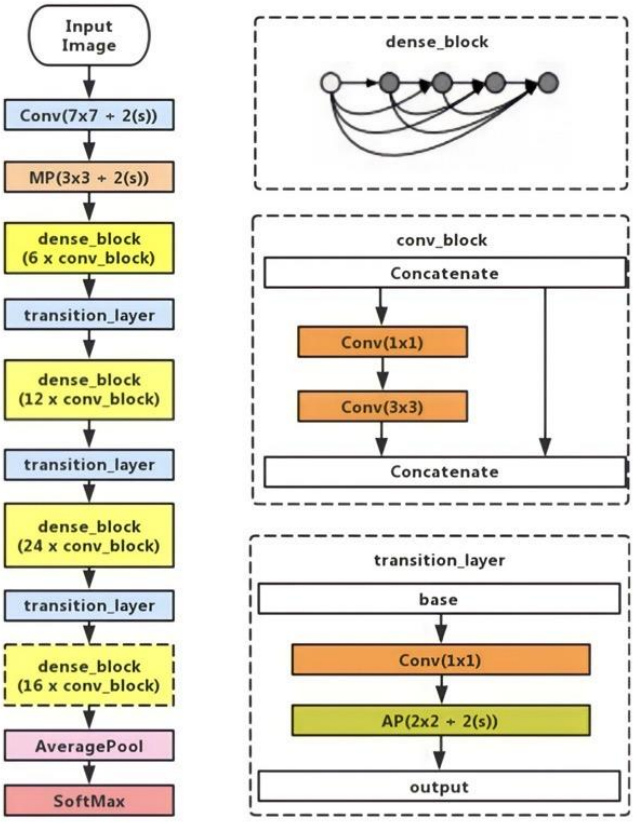


Figura 3.8. Arquitectura DenseNet-121 (izquierda) y bloque denso, bloque convolucional y capa de transición (derecha) (Fuente: Zhang et al., 2021).

detección de la presencia de patología en imágenes de fondo de ojo, es una etapa previa que evitaría el procesado posterior en los casos en los que no se detectan signos patológicos.

3.6.1. Arquitectura empleada

Tal y como se ha comentado previamente, la arquitectura CNN que se decidió emplear para este trabajo fue DenseNet-121, debido a los buenos resultados que se consiguieron alcanzar para la tarea de cribado automático de patologías en nuestro estudio previo (Muñoz Zamarro, 2020). Otra característica de esta arquitectura es que es capaz de evaluar cada imagen de entrada en muy poco tiempo (aproximadamente 8 segundos). Luego, esto permite que la red tarde menos en culminar su entrenamiento y sea capaz de analizar rápidamente las retinografías empleadas. No obstante, han sido necesarias algunas modificaciones para adaptar el modelo a nuestra tarea específica. La estructura final se muestra en la Figura 3.9.

En primer lugar, se partió de la arquitectura base DenseNet-121, tomando como entradas las imágenes preprocesadas de la BD utilizada. La dimensión de dichas imágenes era de 224×244×3 y el espacio de color RGB (de ahí que su profundidad sea igual a 3 por los tres canales de color). Como tipo de agrupación, se ha utilizado *average pooling* y, tras el modelo base, se han añadido tres capas completamente conectadas al final de la red. La primera de ellas, formada por 1024 neuronas y la segunda, formada por 512 neuronas, contaban con una función de activación ReLU, que conduce a una convergencia más rápida en comparación con las no linealidades de la tangente hiperbólica o de la sigmoide. Esta combinación de funciones de activación ya ha sido utilizada con éxito en estudios previos (Mittal & Bhatnagar, 2021). La tercera capa, de salida, tenía 2 neuronas y la función de activación *softmax*, a diferencia de nuestro antiguo trabajo donde la última capa tenía 1 neurona y la función de activación sigmoide (Muñoz Zamarro, 2020). De esta manera, ambas neuronas se corresponden con las dos clases a discriminar: imágenes sanas e

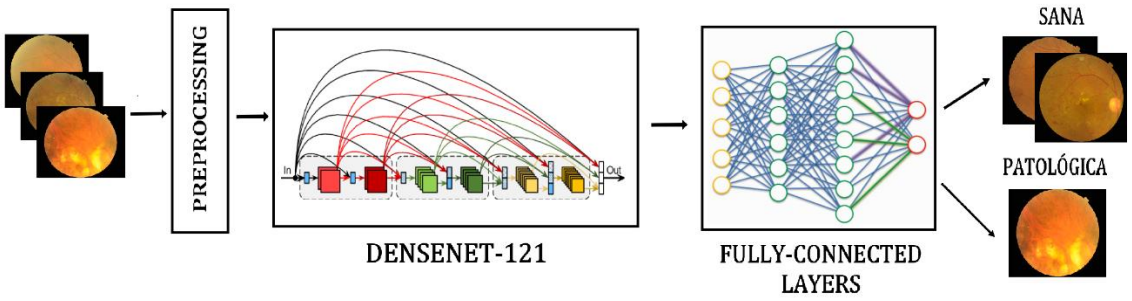


Figura 3.9. Estructura final de la red empleada para el cribado de patología en retinografías.

imágenes patológicas. Aunque supone un cambio menor a nivel de arquitectura, ha sido necesario para la posterior implementación de XAI sobre nuestro modelo. La razón de ello se debe a que la librería *software* (*iNNvestigate*) que se ha utilizado, no admite modelos con otro tipo de función en la capa final de la red. La función de activación *softmax* se emplea en problemas de clasificación multiclase y devuelve una distribución de probabilidades entre las salidas. Es decir, cada una de las salidas da cuenta de la probabilidad de que la imagen de entrada pertenezca a la clase correspondiente (Aggarwal, 2018).

Para la implementación de todos los métodos XAI (excepto SHAP), se siguió el proceso siguiente partiendo del modelo ya entrenado. Primero, se eliminó la última capa de la red (de tipo *softmax*) para crear cada tipo de analizador. Seguidamente, haciendo uso del mismo, se analizó cada una de las imágenes utilizadas. Finalmente, se aplicaron diferentes métodos de postprocesado, para obtener el mapa de atribución adecuado en cada caso. En el caso de SHAP, los gráficos de salida se obtuvieron directamente haciendo uso del algoritmo *Deep Explainer* (SHAP documentation, 2018), seguido también del postprocesado correspondiente.

La función de coste se emplea para calcular el error entre las etiquetas reales y las predicciones estimadas por el modelo. En base a este error, es posible optimizar los parámetros de la red. En nuestro modelo, se ha utilizado la entropía cruzada categórica (*categorical cross entropy*) al utilizar 2 neuronas en la capa de salida. Su expresión matemática para cada imagen, se recoge en la ecuación (3.10) (Liu et al., 2018).

$$L_{CE\,j} = -\sum_{i=1}^{C} t_i * \log(y_i) \tag{3.10}$$

En ella, C denota el número de categorías (en nuestro caso, es igual a 2), t_i se corresponde con la distribución de probabilidad real (es decir, es igual a 1 si pertenece a la etiqueta verdadera, sino es 0) y finalmente, y_i es la distribución de probabilidad de la predicción. Después, se lleva a cabo una media de todos los índices de error de cada una de las imágenes, para así calcular el valor de error o de pérdida de todo el conjunto de imágenes de entrenamiento (N). Por tanto, la expresión final de esta función de coste se muestra a continuación (ecuación 3.11).

$$L_{CE} = \frac{1}{N}\sum_{j=1}^{N} L_{CE\,j} \tag{3.11}$$

Para la actualización de los parámetros de la red, se ha empleado el optimizador SGD (*Stochastic Gradient Descent*) con un *learning rate* igual a 0.005 y un momento igual a 0.9. Los valores de ambos parámetros se han escogido de acuerdo con los empleados en estudios previos (Doshi et al., 2017; Hua et al., 2020). El SGD se trata de una variante del método típico de descenso del gradiente. La diferencia entre ambos se encuentra en que este último realiza los cálculos sobre todo el conjunto de entrenamiento (proceso redundante e ineficiente), mientras que el SGD solo los calcula sobre una selección aleatoria de datos. Todo ello sumado a que, cuando la tasa de aprendizaje es baja, ambos consiguen el mismo rendimiento. Además, aunque existen otros optimizadores, tales como *Adam* o *Adagrad*, se ha decidido utilizar este puesto que se ha demostrado que es el que mejor generaliza en comparación con el resto (Keskar & Socher, 2017). También, conviene señalar que para lidiar con el desbalanceo existente en la BD empleada, se han ponderado las pérdidas mediante la introducción de un factor sobre la clase menos balanceada, durante el proceso de entrenamiento.

Por último, se explican los hiperparámetros (variables que determinan la estructura y el entrenamiento de la red) que se seleccionaron para el modelo desarrollado. Se deben definir antes de entrenar el modelo y son los siguientes (Brownlee, 2018; Smith, 2018):

- **Tamaño del *batch*** (*batch size*). Define el número de imágenes que pasan por la red para ser clasificadas antes de actualizar sus parámetros. En general, tiene sentido escoger un *batch size* lo más grande posible dada la arquitectura de la red y el tamaño de la imagen. Cuanto mayor sea su valor, más rápido se efectuará el entrenamiento de la red. Sin embargo, si éste es demasiado grande, entonces se necesitarán más épocas de entrenamiento para alcanzar un buen rendimiento. En nuestro caso, se ha escogido su valor de manera que, en cada época, se utilizaran una sola vez todas las imágenes del conjunto de entrenamiento y se evaluaran, a continuación, todas las imágenes del conjunto de validación con pasos de idéntico tamaño de *batch*. Es decir, tratando de cumplir que el tamaño de *batch* sea divisor del conjunto de imágenes a procesar. En base a esto, se ha decidido escoger un valor de batch size igual a 16 (ya que el número total de imágenes de entrenamiento era 1200). En el conjunto de validación, se ha utilizado un valor igual a 10, puesto que el número total de imágenes era 300.

- **Número de épocas** (*epochs*). Define el número de veces que el algoritmo de aprendizaje procesa todo el conjunto de entrenamiento. Su valor se ha fijado a 100 y se ha elegido de manera experimental. Es importante elegir el valor

óptimo de este hiperparámetro ya que, si el número de épocas es demasiado bajo, el entrenamiento finalizará antes de que el modelo converja. Por el contrario, si se fija un número de épocas demasiado elevado, la red necesitará mucho tiempo para entrenarse y podría aparecer el problema del sobreentrenamiento. Esto último consiste en que el modelo se ajusta demasiado a los datos de entrenamiento y no funciona bien cuando se utiliza con otros datos independientes.

- **Tasa de aprendizaje** (*learning rate*). Este es un hiperparámetro de gran relevancia puesto que es el que controla cuánto hay que modificar el modelo en base al error estimado calculado cada vez que se actualizan los pesos. Decidir un valor correcto es una tarea compleja ya que un valor demasiado pequeño podría dar lugar a un proceso largo de entrenamiento que podría fallar y, sin embargo, un valor demasiado grande, podría provocar que el aprendizaje de la red fuese demasiado rápido y el conjunto de pesos no fuera el óptimo. Por esta razón, en nuestro caso se ha inicializado de manera experimental (su valor se ha especificado ya anteriormente) y luego se ha hecho uso de la técnica *ReduceLROnPlateau*. Esta permite reducir el *learning rate* cuando se detecta que el aprendizaje se ha estancado puesto que no hay cambios durante cierto número de épocas. Concretamente, en nuestro modelo se monitoriza el error de validación durante 3 épocas y, si no se observa ninguna mejora, la tasa de aprendizaje se reduce a la mitad de su valor.

Respecto a XAI, en la implementación de algunos métodos, también fue necesario ajustar el valor de algunos hiperparámetros. No obstante, cada uno de ellos se explica con mayor detalle en un apartado posterior.

3.6.2. Técnicas de DL para optimización de la red

Con el objetivo de optimizar el funcionamiento de la CNN empleada, se han aplicado cuatro técnicas de DL, que se pasan a explicar a continuación. La primera de ellas, denominada *data augmentation*, adopta el enfoque de ampliar artificialmente el conjunto de datos (a partir de los disponibles) con los que las redes neuronales profundas pueden entrenar. En el caso de las imágenes, se aplican varias transformaciones aleatorias que generan nuevas imágenes también válidas. De esta forma, se consigue que el modelo no se entrene en cada época exactamente con las mismas imágenes, sino que cada vez irá aprendiendo de un conjunto de imágenes diferente. Esta técnica ha demostrado su eficacia puesto que permite obtener mejores resultados ya que se consigue que el sistema desarrollado generalice mejor, a la misma vez que se produce un aumento en su capacidad de aprendizaje y una

disminución en el problema del sobreajuste (exceso de adaptación del modelo al conjunto de entrenamiento). No obstante, hay que tener en cuenta que, si se eligen transformaciones que producen imágenes irreales, esto provocaría un efecto negativo en el entrenamiento de la red puesto que ésta no aprendería las características de las imágenes válidas (Aggarwal, 2018).

Teniendo esto en consideración, las transformaciones que se aplicaron a nuestro modelo fueron de dos tipos: rotación aleatoria y volteo. Mediante la rotación aleatoria, la imagen se gira un número de grados. Con el volteo, la imagen se refleja o bien respecto al eje vertical o bien, respecto al horizontal. Concretamente, en nuestro modelo se aplicó una rotación aleatoria con un rango de 0^{o} a 50^{o} y un volteo tanto horizontal como vertical de las imágenes, siguiendo el trabajo de (Xu et al., 2017).

La segunda técnica de DL que se ha aplicado ha sido el aprendizaje de transferencia o *transfer learning*. Se trata de una de las técnicas más utilizadas en estos sistemas puesto que evita tener que entrenar una red desde cero, al partir de una preentrenada. Es decir, las redes se entrenan previamente con grandes conjuntos de datos y posteriormente, se ajustan a una cantidad más limitada dependiendo de cuál sea la aplicación concreta. Por tanto, evita tener que disponer de un gran conjunto de imágenes, así como el gran tiempo de computación necesario para efectuar su entrenamiento (Rawat & Wang, 2017). ImageNet es el conjunto de datos más popular para utilizar en el aprendizaje de transferencia. Se trata de una base de datos de gran tamaño ya que alberga más de 15 millones de imágenes etiquetadas de alta resolución que pertenecen aproximadamente a 22.000 categorías (Deng et al., 2010). Los pesos de esta red (ya entrenada para otro problema) se utilizarán para entrenar el sistema desarrollado evitando tener que inicializarlos a cero.

Para implementar esta técnica, existen diferentes estrategias. En nuestro caso concreto, se ha partido de un modelo base preentrenado y se ha eliminado la última capa *fully-connected.* Posteriormente, se han añadido tres capas totalmente conectadas para adaptar la arquitectura a nuestra tarea específica. La razón por la que se ha decidido aplicar *transfer learning* en nuestro modelo, se debe a su éxito en otros estudios existentes en los que se demuestra que con ella, la precisión de los resultados aumenta a la vez que contribuye también a reducir significativamente el sobreajuste de la red (Wan et al., 2018; Xu et al., 2019).

Asociada al concepto del *transfer learning*, la tercera técnica adoptada fue el ajuste fino o *fine-tuning*. Las primeras capas de las CNN son las encargadas de detectar las características generales de bajo nivel de las imágenes tales como son los patrones o

bordes. Sin embargo, las características más relevantes para el problema bajo estudio son identificadas por las capas finales de la red. Esta técnica consigue un ajuste más fino de los parámetros de la red reentrenando algunas o todas las capas de una CNN empleando imágenes específicas del problema concreto. Al igual que el resto de las técnicas, existen diferentes alternativas para su implementación como por ejemplo, congelar todas las capas del modelo base convolucional y reentrenar solo las capas completamente conectadas (Saeed et al., 2021). No obstante, en nuestro modelo, se decidió realizar el reentrenamiento de todas las capas ya preentrenadas puesto que fue con la que mejores resultados se obtuvieron para nuestra tarea concreta.

Por último, se decidió aplicar también una técnica de regularización muy utilizada conocida como *dropout*, con el objetivo de evitar el sobreentrenamiento del modelo. Esta técnica funciona eliminando probabilísticamente (es decir, de manera aleatoria) algunas de las neuronas de la red a medida que se va entrenando, de tal forma que aquellas que son canceladas vuelven a aprender evitando así el sobreajuste. Conviene señalar que, al eliminar una neurona, se eliminan también sus conexiones de entrada y de salida asociadas. Por tanto, esta técnica tiene el efecto de simular un gran número de redes con una estructura muy diferente y, a su vez, hacer que los nodos de la red sean generalmente más robustos a las entradas (Nguyen et al., 2020).

El *dropout* se implementa manteniendo una neurona activa con cierta probabilidad que se mantiene constante a lo largo de la fase de entrenamiento. Es decir, por ejemplo, si se aplicase una probabilidad igual al 0.5 entre la capa 7 y 8 del modelo, esto querría decir que todos los valores que van a salir de la capa 7 tienen solo un 50% de probabilidad de llegar a la capa 8. Al igual que en otros estudios previos, en nuestro trabajo, se decidió emplear un índice de *dropout* igual a 0.25 (Minarno et al., 2022).

3.7. Métodos XAI

Para la toma de decisiones, resulta imprescindible que tanto los profesionales del aprendizaje automático como los usuarios finales observen las características relevantes que emplea un sistema de IA. La explicación de las decisiones de diagnóstico y tratamiento a todas las partes implicadas es una parte integral del sistema sanitario moderno. Los retos éticos y legales de este ámbito exigen que las decisiones sean más explicables, transparentes y comprensibles para los usuarios. Todo esto ha dado lugar al avance en el desarrollo de sistemas XAI para el

diagnóstico médico (Singh et al., 2020a).

Tal y como ya se comentó en el capítulo de introducción, para este trabajo se han implementado varios métodos basados en la atribución, que tienen por objetivo determinar la contribución de una característica de entrada en la neurona de salida de la clase correcta en un problema de clasificación. Para la disposición de dichas atribuciones se han empleado mapas de calor, también denominados en este contexto, mapas de atribución (Singh et al., 2020a). Además, para la implementación de estos métodos en Python, se han empleado dos librerías: iNNvestigate (Alber et al., 2019) y SHAP (Lundberg & Lee, 2017) . A continuación, se detallan los diferentes métodos de atribución empleados.

3.7.1. SHAP

El método conocido como SHAP es el enfoque unificado para la interpretación de modelos (Lundberg & Lee, 2017). Este *framework* combina técnicas propuestas en otros métodos desarrollados anteriormente (tales como LIME y DeepLIFT) bajo la clase de atribución de características aditivas y los métodos que pertenecen a esta clase contienen modelos de explicación con una función lineal de variables binarias. Por tanto, proporciona visualizaciones interactivas e intuitivas que muestran qué características tienen mayor importancia para una predicción determinada y para el modelo general (Lundberg & Lee, 2017).

Concretamente, SHAP asigna a cada característica un valor de relevancia para una predicción concreta e incluye resultados teóricos que demuestran que existe una solución única con un conjunto de propiedades deseables. Si consideramos al modelo real como f, para explicarlo se necesita un modelo de explicación g. Dada una instancia única $\mathbf{z}$, para su explicación, g utiliza un vector de características simplificadas o vector de coalición $\mathbf{z}'$ con una función de mapeo h, tal que $\mathbf{z} = h(\mathbf{z}')$. SHAP hace uso del concepto de valores *Shapley* para explicar la contribución de cada característica en el resultado del modelo. Además, en (Lundberg & Lee, 2017) se explicaron tres propiedades deseables de los métodos clásicos de estimación de dichos valores. Estas propiedades eran la precisión local, la ausencia de datos y la consistencia. Dado que calcular los valores *Shapley* exactos suponía un reto, los autores decidieron aproximar estos valores bajo los métodos de atribución de características aditivas. En este contexto, se definieron estos valores como $\boldsymbol{\phi} \in \mathbb{R}$, de manera que ϕ_i hacía referencia a la atribución de característica para la característica i y su definición matemática es la siguiente (ecuación 3.12):

$$\phi_i = \sum_{S \subseteq F\{i\}} \frac{|S|!\, F(|F| - S - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \qquad (3.12)$$

En ella, S es un subconjunto de F, que representa todos los conjuntos de características. El modelo entrenado con S y la i-ésima característica, se denota como $f_{S \cup \{i\}}$, el modelo entrenado sin esa característica como f_S y finalmente, x_S representa los valores de las características en el conjunto S. También, resulta necesario definir un tamaño máximo de coalición M, de manera que $\mathbf{z}' \in \{0,1\}^M$. El 0 significa que el valor de la característica está ausente y el 1 que está presente. La función matemática completa del modelo de explicación correspondiente a SHAP se define en la siguiente ecuación (3.13).

$$g(\mathbf{z}') = \phi_0 + \sum_{i=1}^{M} \phi_i \, \mathrm{z}'_i \qquad (3.13)$$

Teniendo en cuenta lo anterior, en este trabajo se ha empleado el método SHAP para detectar los píxeles de las imágenes de fondo de ojo que más influyen en la clasificación realizada por la CNN empleada en el algoritmo.

3.7.2. *Input x Gradient*

Este método de atribución se propuso en un principio como una técnica para mejorar la nitidez de los mapas de atribución obtenidos simplemente mediante el cálculo del gradiente de la neurona de salida con respecto a la entrada. Esta técnica calcula cada valor de atribución (que es la contribución de cada característica de entrada a la red) tomando las derivadas parciales (con signo) de la salida con respecto a la entrada y multiplicándolas por esta última. Su definición matemática se muestra en la siguiente expresión (3.14):

$$R_i{}^c(\mathbf{x}) = x_i \frac{\partial S_i(\mathbf{x})}{\partial x_i} \qquad (3.14)$$

En esta ecuación, $\mathbf{x} = [x_1, x_2, x_3, \dots, x_N] \in \mathbb{R}^N$ hace referencia a la señal de entrada y $\mathbf{S}(\mathbf{x}) = [S_1(x), S_2(x), S_3(x), \dots, S_C(x)]$ a la señal de salida de la red, siendo C el número total de neuronas de salida. Finalmente, dada una neurona objetivo específica (c), se calcula la contribución $\mathbf{R^c} = [R^c{}_1, R^c{}_2, R^c{}_3, \dots, R^c{}_N] \in \mathbb{R}^N$ de cada característica de entrada x_i (Ancona et al., 2017).

Este método se basa en la idea de que el gradiente nos indica la importancia de una dimensión y la entrada nos permite saber con qué intensidad se expresa esa dimensión en la imagen. Al complementar ambos, la atribución de una dimensión será elevada solo si la dimensión parece ser relevante para el resultado y su valor también es alto. No obstante, esta técnica XAI presenta ciertos problemas puesto que el gradiente nos proporciona información local cuando los cambios en la entrada son pequeños. Por tanto, al expresar funciones complejas que generan cambios bruscos podría ocurrir que el gradiente no funcionase correctamente y nos proporcionara información incorrecta. Con el objetivo de minimizar este problema, se desarrollaron otros métodos XAI tales como *SmoothGrad* e IG los cuales se pasarán a explicar a continuación.

3.7.3. LRP

El método conocido como LRP (*Layer wise relevance propagation*) es una técnica XAI que permite explicar las predicciones individuales de las redes neuronales en términos de la variable de entrada. Dada una entrada y la predicción correspondiente, asigna una puntuación a cada una de las variables de entrada indicando en qué medida han contribuido a dicha predicción de la red. Concretamente, LRP funciona mediante la propagación inversa de la predicción ($f(x)$) por medio de reglas heurísticas de propagación que se aplican a cada capa de la red. Este procedimiento de propagación está sujeto a una propiedad de conservación donde la cantidad neta, o relevancia, recibida por cualquier neurona de la capa superior se redistribuye en la misma cantidad a las neuronas de la capa inferior (de manera análoga a las leyes de conservación de Kirchoff en los circuitos eléctricos) (Bach et al., 2015).

Sean j y k dos neuronas que se encuentran en capas consecutivas de una red neuronal. La propagación de las puntuaciones de relevancia $(R_k)_k$ en una capa determinada a las neuronas de la capa inferior, se consigue aplicando la regla que se muestra en la ecuación siguiente (3.15):

$$R_j = \sum_k \frac{z_{jk}}{\sum_j z_{jk}} R_k \tag{3.15}$$

donde la cantidad z_{jk} modela la medida en la que la neurona j ha contribuido a que la neurona k sea relevante (Bach et al., 2015). El procedimiento de propagación finaliza cuando se alcanzan las características de entrada. En la regla anterior, el

denominador sirve para hacer cumplir la propiedad de conservación, ya comentada. Además, si se emplea dicha regla para todas las neuronas de la red, se puede verificar fácilmente la propiedad anterior por capas $\sum_j R_j = \sum_k R_k$ y, por extensión, la propiedad de conservación global $\sum_i R_i = f(x)$. Aunque el LRP difiere claramente del enfoque de *DeepTaylor*, cada paso del procedimiento de propagación puede modelarse como una descomposición de Taylor propia realizada sobre cantidades locales en el mapa (Bach et al., 2015).

Entre las principales reglas de propagación se encuentran la regla básica (LRP-0), la regla gamma (LRP$-\gamma$) y la regla épsilon (LRP$-\epsilon$). Respecto a esta última, ya se ha utilizado con éxito en estudios previos de tareas de clasificación de imágenes médicas mediante el uso de CNNs (Bourdon et al., 2021; Nedjar et al., 2022). Consiste en añadir un pequeño término positivo ϵ en el denominador de la expresión para calcular las R_j (ver expresión 3.16). El papel de este término es absorber cierta relevancia cuando las contribuciones a la activación de la neurona k son débiles o contradictorias. Luego, a medida que aumenta, solo los factores explicativos más destacados sobreviven a la absorción. Por tanto, esto suele conducir a explicaciones más dispersas en términos de las características de entrada y menos ruidosas (Bach et al., 2015). Para este trabajo, se ha decidido implementar LRP$-\epsilon$ con diferentes valores de dicho término para así poder comprobar las diferencias.

$$R_j = \sum_k \frac{z_{jk}}{\epsilon + \sum_j z_{jk}} R_k \qquad (3.16)$$

3.7.4. *Integrated Gradients* (IG)

Se trata de un método de atribución XAI que combina la invariabilidad de implementación de los gradientes junto con la sensibilidad de técnicas tales como LRP o *DeepLIFT*. Entre sus principales ventajas se encuentra su sencillez de implementación puesto que no requiere ninguna modificación de la red neuronal original, sino que solo se necesita utilizar varias veces el operador de gradiente estándar (Sundararajan et al., 2017).

En el estudio donde se desarrolló IG, se especificaron dos axiomas (o características deseables) que debía cumplir todo método de atribución: sensibilidad e invariabilidad de la implementación. El primero de ellos se satisfacía si para cada entrada y línea de base (línea de referencia donde la predicción es neutral, por ejemplo, en las CNN de reconocimiento de objetos, suele ser la imagen negra) que difirieran en una característica, pero tenían predicciones diferentes, entonces la

característica diferente debía recibir una atribución no nula. El segundo axioma se cumplía si las atribuciones eran siempre idénticas para dos redes funcionalmente equivalentes (sus resultados son iguales para todas las entradas, aunque sus implementaciones fuesen muy diferentes). Como ya se ha comentado previamente, en ciertas situaciones los métodos que calculan el gradiente o éste multiplicado por la señal de entrada para obtener cada valor de atribución, no obtienen los valores adecuados y se centran en rasgos irrelevantes. En base a lo anterior, se dice que estos métodos rompen la sensibilidad puesto que la función de predicción puede aplanarse en la entrada y, por tanto, tener un gradiente cero a pesar de que el valor de la función en la entrada sea diferente al de la línea base (Sundararajan et al., 2017).

Formalmente, supongamos que tenemos una función $F: \in \mathbb{R}^N \rightarrow [0,1]$ que representa una CNN. La señal de entrada se representa como $\mathbf{x}: \in \mathbb{R}^N$ y la línea de base como $\mathbf{x}': \in \mathbb{R}^N$ y consideramos como el camino recto (en $\mathbb{R}^N$) aquel que va desde $\mathbf{x}'$ hasta x. Con este método se calcularán los gradientes en todos los puntos del camino que se acumularán para obtener los gradientes integrados (de ahí su nombre). Concretamente, estos últimos se definen como la integral de los gradientes a lo largo del camino recto. Luego, por tanto, el gradiente integrado a lo largo de la *i*-ésima dimensión se define con la expresión matemática de la ecuación (3.17):

$$IG_i(\mathbf{x}) = (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial \mathrm{F}\left(\mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}')\right)}{\partial x_i} d\alpha \qquad (3.17)$$

En la práctica, la integral de la ecuación anterior puede aproximarse eficazmente mediante una suma. Simplemente se suman los gradientes en los puntos que ocurren en intervalos suficientemente pequeños a lo largo de la trayectoria en línea recta desde la línea base $\mathbf{x}'$ hasta la entrada $\mathbf{x}$. Esta aproximación se muestra en la ecuación ((3.18):

$$IG_i{}^{aprox}(x) = (x_i - x_i') \sum_{k=1}^{m} \frac{\partial F(x' + \frac{k}{m}(x - x'))}{\partial x_i} \frac{1}{m} \qquad (3.18)$$

donde m representa el número de pasos en la aproximación de *Riemman* de la integral. En (Sundararajan et al., 2017), se especifica que entre 20 y 300 pasos son suficientes para aproximarse a la integral (dentro del 5%). Por ello, en este trabajo, se han realizado diferentes pruebas variando este último parámetro para comprobar como difieren los resultados.

3.7.5. *SmoothGrad*

Este método de atribución permite reducir el ruido visual y puede combinarse con otros algoritmos de mapas de atribución (o de sensibilidad). La idea central en la que se basa consiste en tomar la imagen de interés, muestrear imágenes similares añadiendo ruido a la imagen original y, finalmente, calcular la media de los mapas de sensibilidad resultantes de cada imagen muestreada. Para lograrlo, se basa en los métodos que emplean el cálculo del gradiente de la neurona de salida con respecto a la entrada para así obtener cada valor de atribución. No obstante, presenta algunas diferencias ya que, al igual que el método IG, pretende minimizar el problema de que el gradiente puede fluctuar bruscamente a pequeñas escalas. En otras palabras, el ruido aparente que se ve en un mapa de sensibilidad podría deberse a variaciones locales sin sentido en las derivadas parciales (Smilkov et al., 2017).

Matemáticamente, si denominamos C al conjunto de todas las clases de un sistema de clasificación y S_c a la función de activación para cada clase $c \in C$, el gradiente de S_c (ecuación 3.19) en un punto determinado será menos significativo que una media local de los valores del gradiente. Esto sugiere una nueva forma de crear mapas de atribución mejorados: en lugar de basar la visualización directamente en el gradiente de ∂S_c, se propone un suavizado (*smoothing*) de dicho gradiente empleando un núcleo gaussiano. En un espacio de entrada de alta dimensión, el cálculo directo de esta media local es inviable, por lo que se calcula una simple aproximación estocástica. En concreto, se toman muestras aleatorias en una vecindad de una entrada **x** y se promedian los mapas de sensibilidad resultantes. Esto se traduce en la expresión matemática que se muestra en la ecuación (3.20), en la que n es el número de muestras y $N(0, \sigma^2)$ representa ruido gaussiano con desviación estándar σ (Smilkov et al., 2017).

$$M_c(\mathrm{x}) = \frac{\partial S_c(\mathbf{x})}{\partial \mathbf{x}} \tag{3.19}$$

$$\widehat{M_c}(\mathbf{x}) = \frac{1}{n}\sum_{1}^{n} M_c(\mathbf{x} + N(0, \sigma^2)) \tag{3.20}$$

Luego, por tanto, *Smoothgrad* tiene dos hiperparámetros que se pueden ajustar, que son σ y n. En (Smilkov et al., 2017) se determina que la aplicación de un 10% – 20% de ruido (σ) parece equilibrar la nitidez del mapa de sensibilidad a la vez que se mantiene la estructura de la imagen original. Con respecto al número de muestras (n), en ese mismo estudio se demuestra que, como era de esperar, el gradiente estimado se suaviza a medida que aumenta el tamaño de la muestra. A partir de un

valor igual a 50, los cambios aparentes en las visualizaciones fueron poco significativos. Teniendo en cuenta lo anterior, en este trabajo se han hecho diferentes pruebas en la implementación de este método XAI. Concretamente, se ha establecido el valor de n a 50 y se ha ido variando el nivel de ruido.

3.7.6. *DeepTaylor*

Técnica XAI basada en la descomposición profunda de Taylor, que utiliza eficazmente la estructura de la red retropropagando las explicaciones de la capa de salida a la de entrada. Se inspira en el paradigma de "divide y vencerás" y explota la propiedad que la función aprendida por una red profunda se descompone en un conjunto de subfunciones más sencillas (forzadas estructuralmente por la conectividad de la red neuronal o porque se producen como resultado del entrenamiento). Estas subfunciones pueden aplicarse localmente a subconjuntos de pixeles o pueden operar a un determinado nivel de abstracción en función de la capa de la red en la que se encuentren (Montavon et al., 2017).

Supongamos que la función $f(x)$ codificada por una neurona de salida x_f, se ha descompuesto en el conjunto de neuronas de una capa determinada. x_j será una de esas neuronas, R_j su relevancia asociada y $\{x_i\}$ el conjunto de neuronas de la capa inferior a las que x_j está conectada. La descomposición de $R_j(\{x_i\})$ se puede obtener mediante la descomposición de Taylor. Sin embargo, hay que tener en cuenta que, en la práctica, la función de relevancia puede depender de otras variables adicionales de la red, tales como las relevancias de las neuronas de la capa superior $\{x_k\}$ a las que contribuye x_j. Estas dependencias ascendentes-descendentes incluyen la información necesaria para determinar si una neurona x_j es relevante, no solo en función del patrón que recibe como entrada, sino también en función de su contexto. La descomposición de Taylor de R_j viene dada por la ecuación (3.21).

$$\begin{aligned} R_j &= \left(\frac{\partial R_j}{\partial\{x_i\}}\Big|_{\{\tilde{x}_i\}^{(j)}}\right)^T * \left(\{x_i\} - \{\tilde{x}_i\}^{(j)}\right) + \varepsilon_j \\ &= \sum_i \frac{\partial R_j}{\partial x_i}\Big|_{\{\tilde{x}_i\}^{(j)}} * \left(x_i - \tilde{x}_i^{(j)}\right) + \varepsilon_j \\ &= R_{ij} + \varepsilon_j \end{aligned} \tag{3.21}$$

En ella, se define un punto raíz $\{\tilde{x}_i\}^{(j)}$ que será diferente para cada neurona x_j en la capa actual (de ahí el superíndice j), ε_j denota el residuo de Taylor y $|_{\{\tilde{x}_i\}^{(j)}}$ indica que la derivada se ha evaluado en el punto raíz $\{\tilde{x}_i\}^{(j)}$.El término identificado

como R_{ij} es la relevancia redistribuida de la neurona x_j a la neurona x_i en la capa inferior (Montavon et al., 2017). Para calcular la relevancia total de la neurona x_i, hay que agrupar la relevancia procedente de todas las neuronas $\{x_j\}$ a las que contribuye la neurona x_i, tal y como se muestra en la siguiente ecuación (3.22).

$$R_i = \sum_i R_{ij} \tag{3.22}$$

Finalmente, combinando las dos ecuaciones anteriores se obtiene la ecuación fundamental (3.23) que permite calcular la relevancia total redistribuida sobre la capa anterior.

$$R_i = \sum_i \frac{\partial R_j}{\partial x_i}\Big|_{\{\tilde{x}_i\}^{(j)}} * \left(x_i - \tilde{x}_i^{(j)}\right) \tag{3.23}$$

Para las tareas concretas de clasificación de imágenes, los espacios de píxeles suelen estar sujetos a restricciones en el rango de sus valores, donde una imagen tiene que estar en el dominio $B = \{\{x_i\} : \forall_{i=1}^{d}\, l_i \leq x_i \leq h_i\}$ donde $l_i \leq 0$ y $h_i \geq 0$ son los valores de píxeles más pequeños y más grandes admisibles para cada dimensión. Bajo estas condiciones, se puede restringir la búsqueda de los R_i empleando la denomianda regla z^B (Montavon et al., 2017). En las pruebas llevadas a cabo para este trabajo, se ha implementado *DeepTaylor* limitando los valores de la señal de entrada (*bounded*) y sin limitarlos (*unbounded*), para así comprobar las diferencias entre ambos.

En la Tabla 3.3 se muestra un resumen de todos los métodos de atribución XAI anteriormente comentados, indicando una breve descripción de cada uno de ellos, así como la ecuación matemática general para calcular los valores de atribución.

Método XAI	Descripción	Ecuación matemática
SHAP (Lundberg & Lee, 2017)	Calcula los valores *Shapley* empleando la aproximación de valores bajo los métodos de atribución de características aditivas	$g(\mathbf{z}') = \phi_0 + \sum_{i=1}^{M} \phi_i \, z'_i$
Input x Gradient (Shrikumar et al., 2016)	Inicialmente se propuso como método para mejorar la nitidez de los mapas de atribución y se calcula multiplicando la entrada por la derivada parcial (con signo) de la salida	$R_i{}^c(\mathbf{x}) = x_i \frac{\partial S_i(\mathbf{x})}{\partial x_i}$
LRP (Bach et al., 2015)	Propaga la puntuación de la predicción capa por capa hacia atrás en la red utilizando reglas heurísticas de propagación, tales como la regla ϵ que garantiza que solo los factores explicativos más destacados tienen cierto valor de relevancia	$R_j = \sum_k \frac{z_{jk}}{\sum_j z_{jk}} R_k$
Integrated Gradients (Sundararajan et al., 2017)	Calcula el gradiente medio a medida que la entrada varía desde la línea de base (suele ser cero) hasta el valor de entrada real, a diferencia de Input x Gradient que utiliza una única derivada en la entrada	$IG_i(\mathbf{x}) = (\mathbf{x} - x_i{}') \int_{\alpha=0}^{1} \frac{\partial \mathrm{F}\left(\mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}')\right)}{\partial x_i} d\alpha$
SmoothGrad (Smilkov et al., 2017)	Diseñado para agudizar visualmente las atribuciones producidas por el método del gradiente calcula una versión mejorada promediando el gradiente sobre múltiples entradas con ruido adicional	$\widehat{M_c}(\mathbf{x}) = \frac{1}{n} \sum_{1}^{n} M_c(\mathbf{x} + N(0, \sigma^2))$
DeepTaylor (Montavon et al., 2017)	Encuentra un punto raíz cerca de cada neurona con un valor cercano a la entrada, pero con salida como 0 y lo utiliza para estimar recursivamente la atribución de cada neurona empleando la descomposición de Taylor	$R_i = \sum_i \frac{\partial R_j}{\partial x_i}\Big\|_{\{\tilde{x}_i\}^{(j)}} * \left(x_i - \tilde{x}_i^{(j)}\right)$

Tabla 3.3. Métodos de atribución XAI empleados.

Capítulo 4

Resultados

4.1. Introducción

En el presente trabajo se ha utilizado una CNN con su arquitectura óptima (en base a los resultados de nuestro trabajo previo) para la clasificación de patología en retinografías. Posteriormente, se han aplicado diferentes métodos XAI para explicar y analizar las predicciones realizadas por la red. Además, tal y como ya se ha comentado, se han empleado dos bases de datos distintas.

Tras haber descrito en el capítulo anterior la metodología que se ha seguido para la realización del trabajo, en este capítulo se exponen los resultados obtenidos al aplicar el algoritmo desarrollado. En primer lugar, se explica el modo de evaluación del método detallando las diferentes métricas empleadas. Seguidamente, se muestran los resultados obtenidos durante el entrenamiento y el test. Por último, se presentan los mapas de atribución que se han conseguido con cada una de las técnicas XAI.

4.2. Modo de evaluación

En este trabajo, se ha llevado a cabo un problema de clasificación, por lo que la variable objetivo que se predice solo puede tomar ciertos valores discretos. Como ya se ha comentado, el propósito final del método es clasificar cada imagen de fondo de ojo en dos posibles clases: no patológica o normal (si el sistema no detecta ninguna lesión o afección ocular en la retinografía) y patológica (si el sistema detecta la presencia de algún signo patológico en la imagen de retina). Para ello, se ha utilizado un conjunto de entrenamiento formado por 2720 imágenes (800 de la BD privada y 1920 de la pública), un conjunto de validación de 740 imágenes (100 procedentes de la BD privada y 640 de la pública) y dos conjuntos de test, uno de 100 imágenes (BD privada) y otro de 640 (BD pública).

Si bien la preparación de los datos y el entrenamiento del modelo son pasos clave en el proceso de DL, es igualmente importante analizar su comportamiento y rendimiento para poder optimizarlo. Para ello, es común utilizar ciertas métricas que ofrecen resultados numéricos que permiten dar cuenta de cómo de buena es la CNN implementada (Das & Saha, 2022; Kwasigroch et al., 2018). Concretamente, las métricas que se han empleado son las siguientes: matriz de confusión, precisión, sensibilidad, especificidad y curva ROC. A continuación, se explica más en detalle cada una de ellas.

4.2.1. Matriz de confusión

Se trata de una representación en forma de tabla que permite comparar la clasificación estimada por el modelo automático frente a la clasificación real. Cada columna de la matriz representa el número de retinografías que la red ha predicho como pertenecientes a cada una de las clases, mientras que las filas representan el número de imágenes que había realmente en cada clase. Dado que las clases se enumeran con el mismo orden tanto en las filas como en las columnas, los elementos que han sido correctamente clasificados se ubican en la diagonal de la matriz (Gayathri et al., 2020).

En nuestro caso, únicamente habrá que diferenciar entre las dos clases bajo estudio, que, de cara a la explicación, denominaremos S a la clase correspondiente a imágenes sanas y P a la clase de imágenes con patología. Teniendo esto en cuenta, los elementos de la matriz de confusión serán los siguientes (Tabla 4.1) (Setiawan & Damayanti, 2020):

1. **Verdaderos positivos** (TP, True Positives). Número de imágenes pertenecientes a la clase P que han sido correctamente clasificadas como clase P por el modelo.

2. **Falsos negativos** (FN, False Negatives). Número de imágenes pertenecientes a la clase P que han sido erróneamente clasificadas como clase S por el modelo.

3. **Verdaderos negativos** (TN, True Negatives). Número de imágenes pertenecientes a la clase S que han sido correctamente clasificadas como clase S por el modelo.

4. **Falsos positivos** (FP, False Positives). Número de imágenes pertenecientes a la clase S que han sido erróneamente clasificadas como clase P por el modelo.

		Clasificación real	
		Normales	**Patológicas**
Clasificación estimada	**Normales**	TN	FN
	Patológicas	FP	TP

Tabla 4.1. Matriz de confusión correspondiente al problema de clasificación bajo estudio.

A partir de estos cuatro valores, TP, FN, TN y FP, es posible obtener diferentes estadísticos tales como la precisión, la sensibilidad y la especificidad.

4.2.2. Precisión, sensibilidad y especificidad

La precisión o exactitud (*accuracy*) determina, para una cierta clase, el porcentaje de clasificaciones acertadas respecto al total. En este trabajo, se ha utilizado la precisión categórica cuyo cálculo viene dado por (Baratloo et al., 2015):

$$\text{Precisión} = \frac{\text{TP} + \text{TN}}{TP + TN + FP + FN} \tag{4.1}$$

La sensibilidad (*sensitivity*) proporciona el porcentaje de imágenes patológicas (clase positiva) respecto al total que han sido detectadas de manera adecuada por el sistema. Se calcula de la siguiente manera (Baratloo et al., 2015):

$$\text{Sensibilidad} = \frac{\text{TP}}{TP + FN} \tag{4.2}$$

Por último, la especificidad (*specificity*) proporciona el porcentaje de imágenes sanas (clase negativa) con respecto al conjunto total que el sistema ha sido capaz de detectar correctamente. Su expresión es la siguiente (Baratloo et al., 2015):

$$\text{Especificidad} = \frac{\text{TN}}{TP + FN} \tag{4.3}$$

4.2.3. Curva ROC

La curva ROC (*Receiver Operating Characteristic*) da cuenta del rendimiento del modelo mediante la representación de la sensibilidad (tasa de verdaderos positivos) frente a la especificidad (tasa de falsos positivos), tal y como se muestra en la Figura 4.1. La información bajo esta curva se puede calcular mediante un único parámetro denominado AUC (*Area Under Curve*). De tal forma que, un valor de AUC igual a 0.5 significaría que el modelo efectúa predicciones aleatorias, sin ser capaz de discriminar entre las distintas clases. Por el contrario, un valor igual a 1 supondría que el sistema puede discriminar sin fallos las diferentes clases del problema (este es el caso de un clasificador perfecto). Luego, por lo tanto, cuanto mayor sea este valor, más óptimo será el modelo construido y mayor será su rendimiento (Alcalá et al., 2020).

4.3. Medida de resultados

4.3.1 Fase de entrenamiento

Esta fase tiene por objetivo ajustar los parámetros internos del modelo para su adaptación al problema dado. Para ello, se utiliza una fuente de datos etiquetados y un algoritmo de aprendizaje. Concretamente, para esta etapa se ha utilizado el conjunto de datos de entrenamiento y el de validación. Ambos conjuntos contienen imágenes de las dos BBDD empleadas, lo que da lugar a un total de 2720 imágenes de entrenamiento y 740 de validación . Entre los problemas más comunes durante

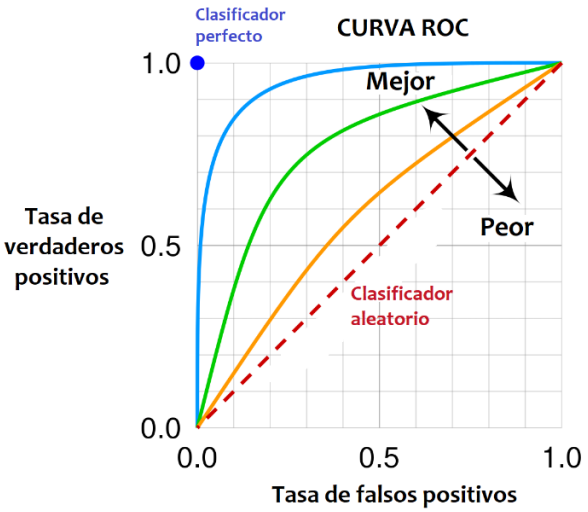


Figura 4.1. Ejemplo de curva ROC para un clasificador perfecto (AUC=1), bueno y malo (Fuente: adaptada de Wikipedia).

esta etapa, destaca el sobreentrenamiento de la red (es decir, el modelo se ajusta demasiado al conjunto de entrenamiento y no es capaz de generalizar cuando se clasifican otras imágenes independientes). Esto último es un aspecto muy importante en los métodos de detección de patología ya que el sistema debe ser capaz de generalizar y, por lo tanto, de clasificar retinografías procedentes de otras BBDD (correspondientes a distintos pacientes).

Respecto al entrenamiento de la CNN, primero se ha observado la evolución de la pérdida y precisión para cada época de entrenamiento (Figura 4.2). Tal y como se puede observar, a medida que se itera la red, las pérdidas van disminuyendo en ambos casos (entrenamiento y validación). No obstante, se comprueba que la pérdida de entrenamiento disminuye en mayor medida respecto a la de validación. La primera llega a ser prácticamente nula (caso ideal) mientras que la segunda se estabiliza en un valor algo más elevado. Por el contrario, con el número de épocas, la precisión va aumentando y de nuevo, se obtienen mejores resultados en el conjunto de entrenamiento. Todo ello se podría deber a que existe algo de sobreentrenamiento en la red.

Posteriormente, se han calculado las diferentes métricas de evaluación para analizar los resultados tanto en el conjunto de entrenamiento como en el de validación. De esta forma, también se podía evaluar si existía o no sobreentrenamiento. La matriz de confusión normalizada para cada caso se muestra en la Figura 4.3. Se comprueba que, en ambos casos, los valores más altos se obtienen en la diagonal de la matriz. Además, en el caso de entrenamiento se obtienen los valores óptimos (los valores de la diagonal son igual a 1 y el resto igual a 0), lo que significa que todas las imágenes han sido clasificadas correctamente por la red. Dicho de otro modo, no

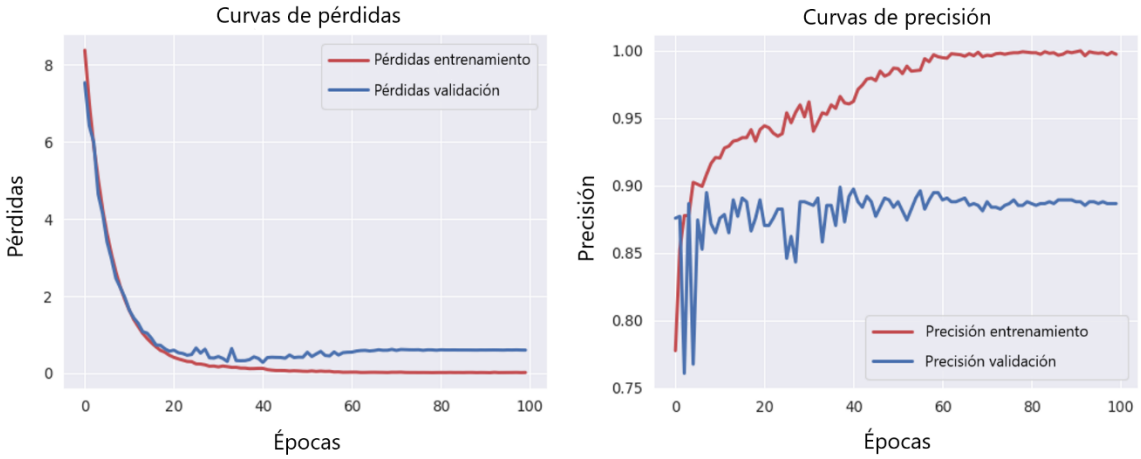


Figura 4.2. (a) Evolución de la pérdida y (b) evolución de la precisión, en función de las épocas de entrenamiento en el conjunto de entrenamiento y de validación, para la arquitectura DenseNet-121

existen casos correspondientes a falsos negativos ni a falsos positivos. En validación, los resultados obtenidos son algo peores, puesto que solo el 86% de los casos se corresponden con verdaderos negativos y el 89% con verdaderos positivos. Como consecuencia, el 11% de las imágenes etiquetadas se corresponden con falsos positivos y el 14% con falsos negativos.

Los valores anteriores se corresponden con una precisión, sensibilidad y especificidad igual al 100% en el conjunto de entrenamiento. En el de validación, se ha obtenido, una precisión igual al 88.65%, una sensibilidad igual al 89.39% y una especificidad del 86.41%.

Por último, se ha calculado la curva ROC y el valor del AUC también para ambos casos, con el objetivo de analizar el aprendizaje del modelo (Figura 4.4). Como era de esperar, el AUC es mayor en el conjunto de entrenamiento puesto que se alcanza el valor óptimo (clasificador perfecto con AUC=1), comparado con el de validación. No obstante, en este último el valor es también muy bueno, puesto que se alcanza un AUC igual a 0.94.

4.3.2 Fase de test

En esta fase, se emplean los parámetros óptimos que se han obtenido en la fase de entrenamiento, para evaluar el modelo final sobre un nuevo conjunto de imágenes de test. Como ya se ha comentado, en este trabajo, se han utilizado dos subconjuntos diferentes para realizar el test, el primero de ellos correspondiente a la BD privada original formado con 100 imágenes y el segundo formado por 640 imágenes procedentes de la BD conocida como RFMiD. Al igual que en la fase anterior, para evaluar el modelo de manera cuantitativa en ambos conjuntos de datos, también se

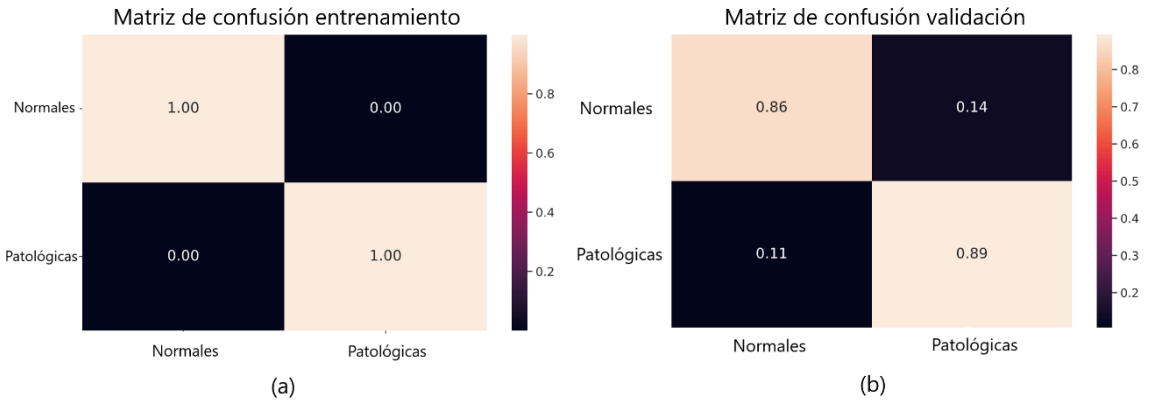


Figura 4.3. Matriz de confusión normalizada para (a) el conjunto de entrenamiento y (b) para el conjunto de validación.

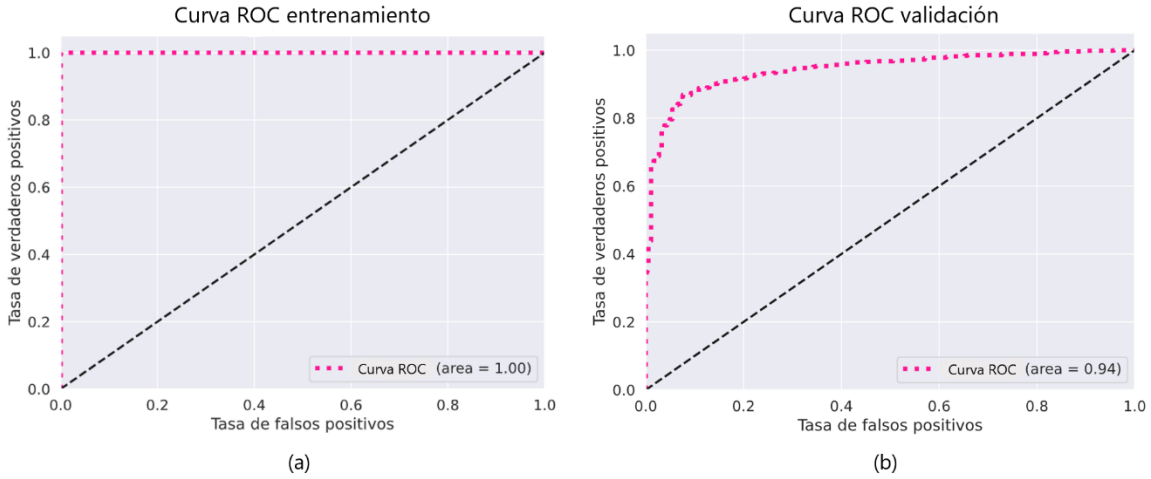


Figura 4.4. Curva ROC para (a) el conjunto de entrenamiento y (b) el conjunto de validación.

ha calculado la matriz de confusión, la precisión, la sensibilidad, la especificidad, la curva ROC y el AUC. Posteriormente, se han implementado seis métodos XAI para interpretar las predicciones realizadas por la red en cada caso. Dado que estos métodos dan como resultados mapas de atribución, se mostrará un ejemplo de cada uno de ellos para el caso de una imagen sana (clase negativa) y de una imagen patológica (clase positiva). También, se incluye, para cada técnica XAI, el tiempo medio que se tarda en analizar una imagen.

La matriz de confusión normalizada para cada conjunto de test se muestra en la Figura 4.5. En ambos casos, se puede apreciar que la diagonal principal de la matriz tiene los valores más altos, por lo que se puede decir que el modelo desarrollado generaliza adecuadamente cuando se utiliza con otras imágenes diferentes a las de entrenamiento. A continuación, en la Tabla 4.2 se recogen los valores calculados con el resto de las métricas. Por último, en la Figura 4.6 se muestra la curva ROC obtenida en cada conjunto de test empleado (el de la BD privada y el de la BD pública).

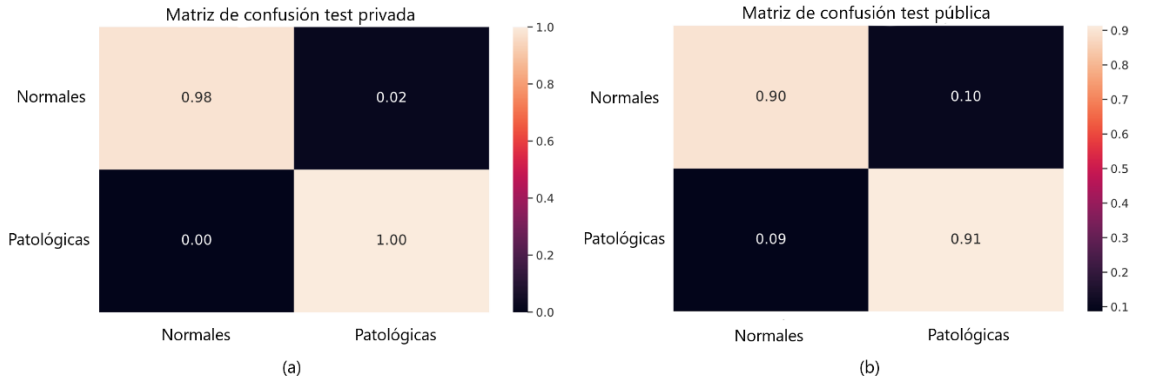


Figura 4.5. Matriz de confusión normalizada para (a) la base de datos privada y (b) pública.

	Precisión (%)	Sensibilidad (%)	Especificidad (%)	AUC
BBDD privada	99.00%	100%	98.00%	0.99
BBDD pública	90.93%	91.30%	89.55%	0.97

Tabla 4.2. Métricas obtenidas en ambos conjuntos de test utilizados.

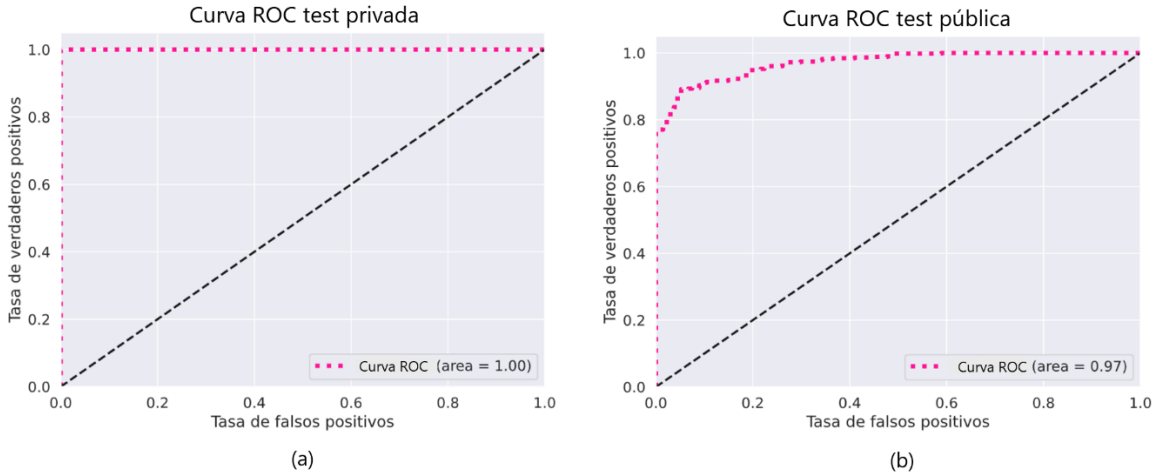


Figura 4.6. Curva ROC para el conjunto de test de (a) la base de datos privada y (b) la base de datos pública.

Para mostrar los resultados obtenidos con las diferentes técnicas XAI se han elegido, a modeo de ejemplo, dos imágenes sanas y dos patológicas procedentes de las dos BBDD que se han empleado en el TFM. Los resultados obtenidos para estas imágenes son representativos de lo que sucede en las restantes imágenes de test. En la Figura 4.7 se muestran las retinografías sanas y patológicas originales de la BD privada y de la BD pública, que se han empleado para visualizar los resultados de XAI.

4.3.2.1. Resultados con SHAP

En la Figura 4.8 se muestran ejemplos de visualización de valores SHAP para una imagen de fondo de ojo sana y otra patológica, procedentes de la BD privada. Para este caso, el tiempo medio de análisis de una imagen ha sido de 11.60 segundos. En cuanto a la BD pública, los ejemplos de valores SHAP tanto para una imagen sin patología como con patología, se muestran en la Figura 4.9. En este caso, se ha obtenido un tiempo medio de análisis muy parecido al anterior, ya que la diferencia entre ambos ha sido solo de medio segundo aproximadamente (11.0058 segundos).

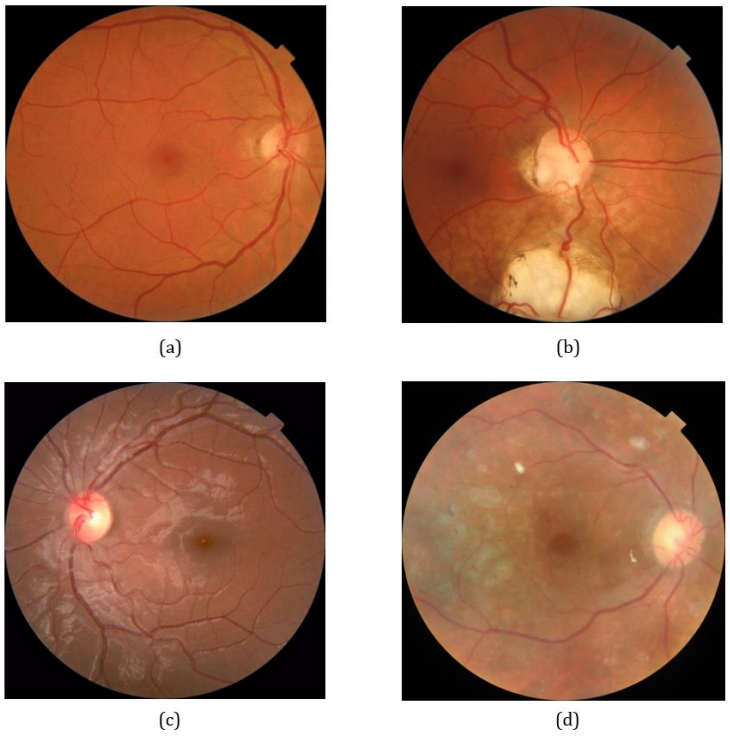


Figura 4.7. Imágenes de fondo de ojo (a) sin patología y (b) con patología procedentes de la base de datos privada y (c) sin patología y (d) con patología procedentes de la base de datos pública.

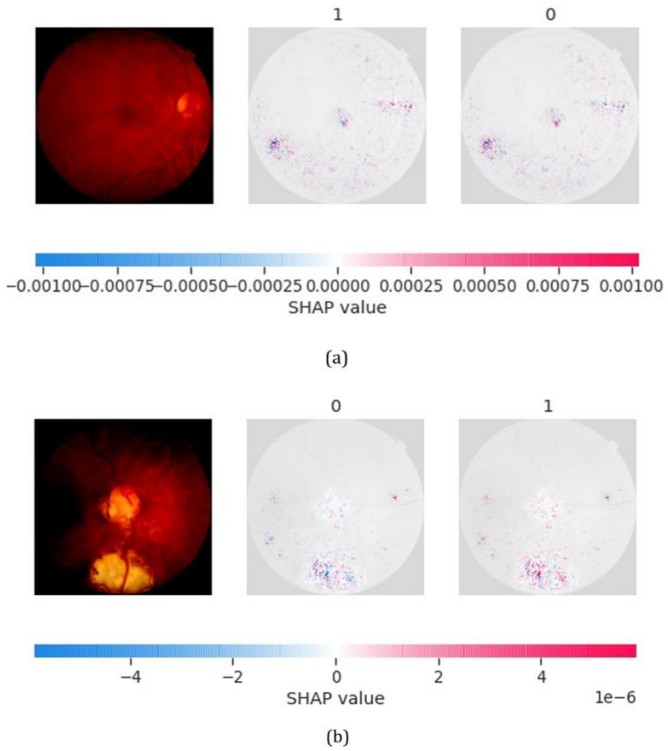


Figura 4.8. Ejemplos de mapas obtenidos con SHAP en la base de datos privada para (a) imagen sin patología y (b) imagen patológica.

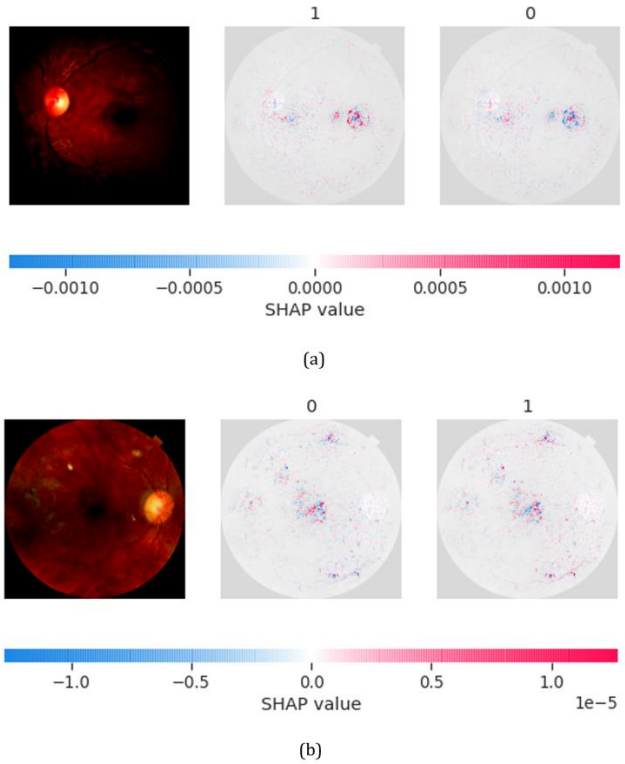


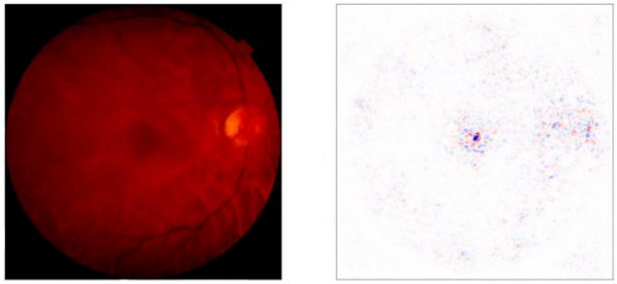
Figura 4.9. Ejemplos de mapas obtenidos con SHAP en la base de datos pública para (a) imagen sin patología y (b) imagen patológica.

4.3.2.2. Resultados con *Input x Gradient*

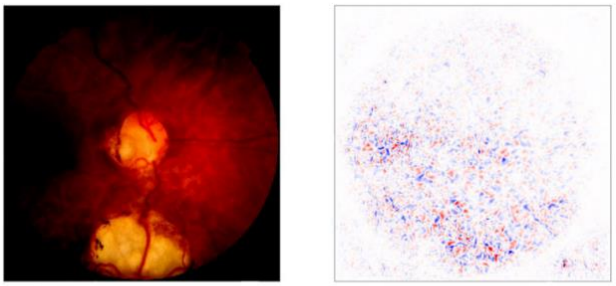
Los mapas de atribución obtenidos con este método, se muestran en la Figura 4.10 (BD privada) y en la Figura 4.11 (BD pública). En el primer caso, se ha obtenido un tiempo medio de análisis de 1.82 segundos. Sin embargo, en las imágenes de la BD pública, este tiempo es algo menor ya que se ha obtenido un valor igual a 1.2252 segundos.

4.3.2.3. Resultados con IG

En este método, tal y como se comentó en el capítulo anterior, se han realizado diferentes pruebas variando el hiperparámetro correspondiente al número de pasos (m) en la aproximación de *Riemman* para calcular la integral de la ecuación que responde a este método. Concretamente, se han obtenido los mapas de atribución para m igual a 64, 100 y 160 pasos. Los dos primeros valores se han elegido teniendo en cuenta otros estudios previos (Alber et al., 2019; Sundararajan et al., 2017). El tercer valor se decidió utilizar ya que es el punto medio del rango recomendado (20-300 pasos).

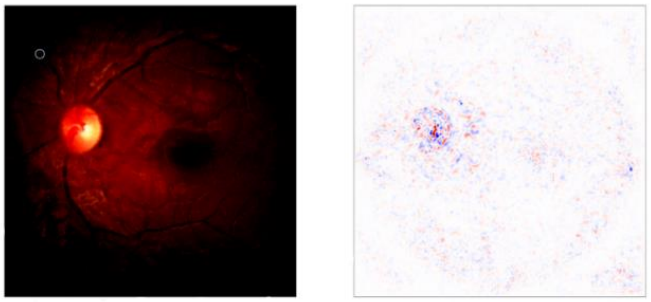


(a)

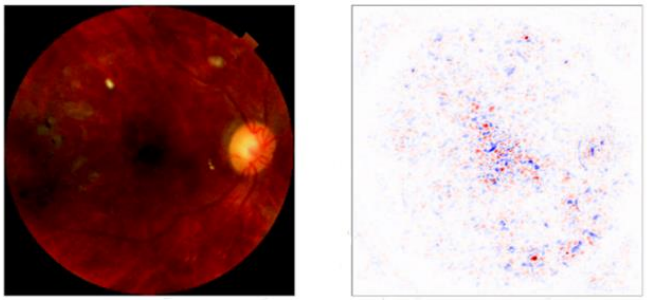


(b)

Figura 4.10. Ejemplos de mapas obtenidos con Input x Gradient en la base de datos pública para (a) imagen sin patología y (b) imagen patológica.



(a)



(b)

Figura 4.11. Ejemplos de mapas obtenidos con Input x Gradient en la base de datos privada para (a) imagen sin patología y (b) imagen patológica.

Los mapas de atribución obtenidos para el caso de $m = 64$ se muestran en la Figura 4.12 (BD privada) y en la Figura 4.13 (BD pública). Aquellos correspondientes a $m = 100$ en la Figura 4.14 (BD privada) y en la Figura 4.15 (BD pública) y finalmente, los mapas asociados a un valor de m igual a 160 se observan en la Figura 4.16 (BD privada) y en la Figura 4.17 (BD pública). En cuanto a los tiempos medios de análisis de cada imagen, todos ellos se especifican en la Tabla 4.3.

4.3.2.4. Resultados con *SmoothGrad*

Al igual que el método anterior, *SmoothGrad* cuenta también con hiperparámetros que se pueden ajustar. Concretamente, son dos: el número de muestras (n) y la desviación estándar del ruido gaussiano (σ). Como ya se comentó, el primer parámetro se decidió fijar a un valor de 50 puesto que, en (Smilkov et al., 2017), se demostró que, a partir de dicho valor, los cambios en las visualizaciones eran poco significativos. Sin embargo, se han obtenido los mapas de atribución variando el valor de σ. En ese mismo estudio se determinó que la aplicación de ruido comprendida entre un 10% - 20%, permitía equilibrar la nitidez de los mapas a la misma vez que mantenía la estructura de la imagen original.

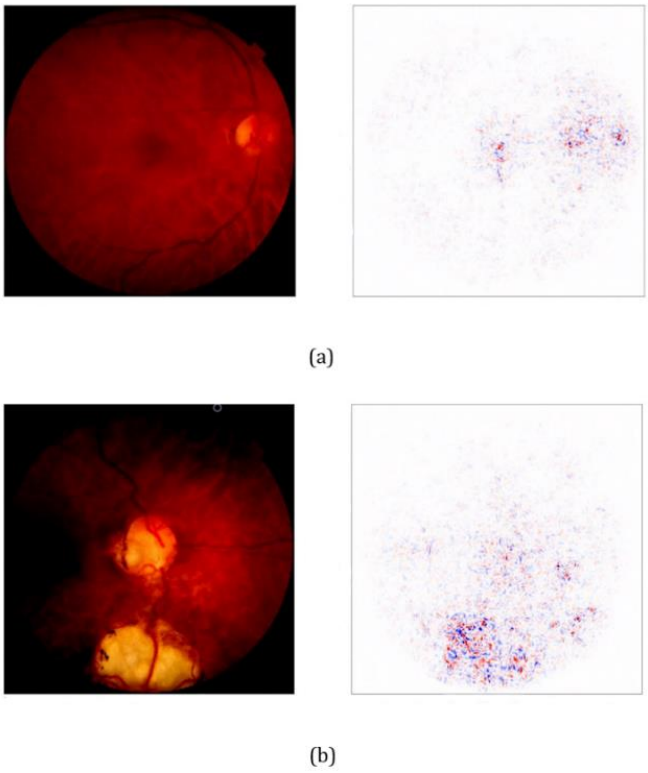
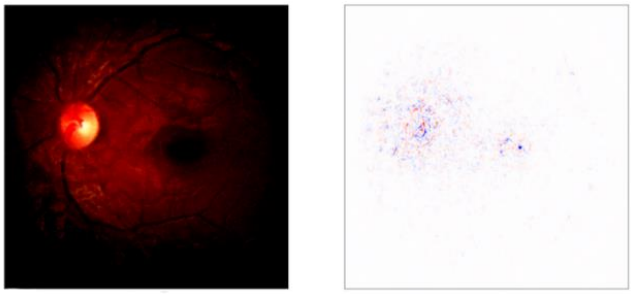
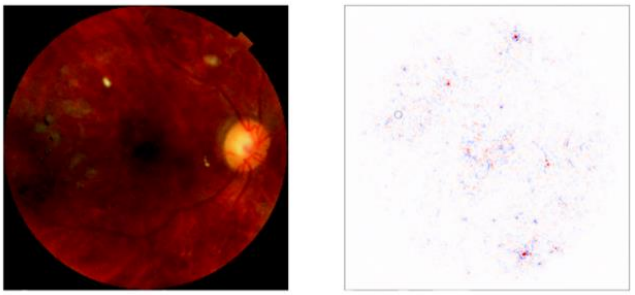


Figura 4.12. Ejemplos de mapas obtenidos con Integrated Gradients (con m=64) en la base de datos privada para (a) imagen sin patología y (b) imagen patológica.

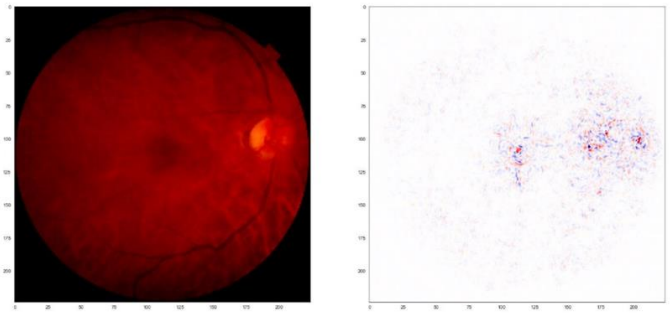


(a)

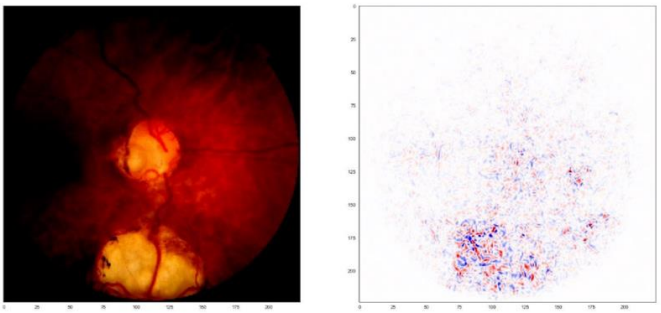


(b)

Figura 4.13. Ejemplos de mapas obtenidos con Integrated Gradients (con m=64) en la base de datos pública para (a) imagen sin patología y (b) imagen patológica.

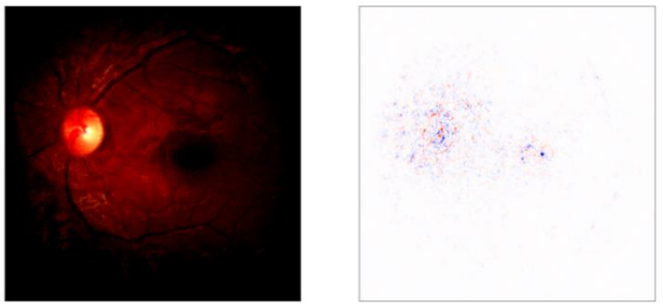


(a)

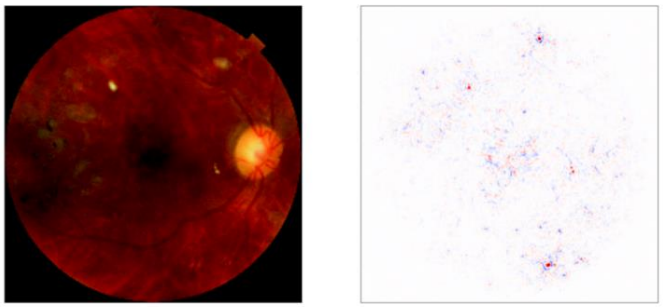


(b)

Figura 4.14. Ejemplos de mapas obtenidos con Integrated Gradients (con m=100) en la base de datos privada para (a) imagen sin patología y (b) imagen patológica.

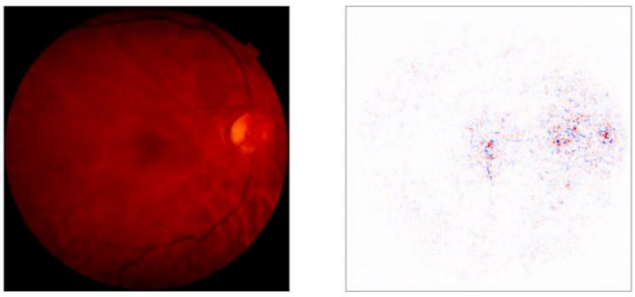


(a)

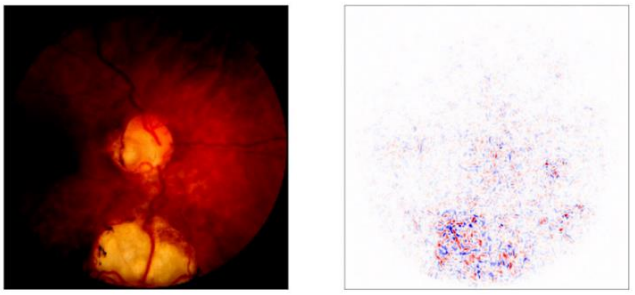


(b)

Figura 4.15. Ejemplos de mapas obtenidos con Integrated Gradients (con m=100) en la base de datos pública para (a) imagen sin patología y (b) imagen patológica.



(a)



(b)

Figura 4.16. Ejemplos de mapas obtenidos con Integrated Gradients (con m=160) en la base de datos privada para (a) imagen sin patología y (b) imagen patológica.

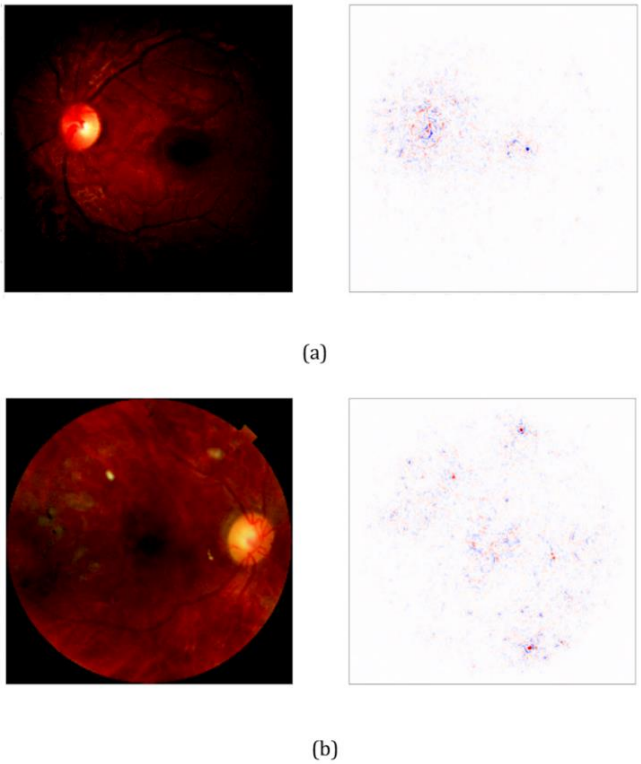


Figura 4.17. Ejemplos de mapas obtenidos con Integrated Gradients (con m=160) en la base de datos pública para (a) imagen sin patología y (b) imagen patológica.

	Pasos (m)	**Tiempo medio de análisis de una imagen (segundos)**
BD privada	64	104.06
	100	198.28
	160	262.16
BD pública	64	100.51
	100	187.42
	160	258.91

Tabla 4.3. Tiempo medio de análisis de una imagen para diferentes valores de m (número de pasos).

Por esta razón, se han realizado tres pruebas distintas, correspondientes a un nivel de ruido (σ) igual a 0.10, 0.15 y 0.20. Respecto al primer caso, los mapas de atribución correspondientes a las imágenes de la BD privada se muestran en la Figura 4.18, mientras que los de la BD pública en la Figura 4.19. Los mapas obtenidos en el segundo caso, se muestran en la Figura 4.20 (BD privada) y en la Figura 4.21

(BD pública). Finalmente, los correspondientes a un valor de $\sigma = 0.20$, son los de la Figura 4.22 y la Figura 4.23, respectivamente. Los tiempos medios de análisis de una imagen se recogen en la Tabla 4.4.

4.3.2.5. Resultados con *DeepTaylor*

Tal y como ya se comentó, en *DeepTaylor* es posible limitar los valores de la señal de entrada (*bounded*) o no limitarlos (*unbounded*). Dado que, en nuestro caso, las imágenes se han normalizado en el rango de valores comprendido entre -1 y 1, los valores de la señal de entrada se han limitado también de igual manera. Para ambos casos, se han obtenido los mapas de atribución. En la Figura 4.24 y en la Figura 4.25, se muestran los correspondientes al caso *bounded* en las imágenes de la BD privada y de la BD pública, respectivamente. Por otro lado, los del caso *unbounded* se muestran en la Figura 4.26 y en la Figura 4.27, también para las imágenes de ambas BBDD. Finalmente, en cuanto a los tiempos medios de análisis, en el caso *bounded* se tarda 1.75 segundos en analizar una imagen procedente de la BD privada y 1.07 segundos en analizar una de la BD pública. Sin embargo, cuando no se limitan los valores de la señal de entrada, el tiempo medio de análisis para la BD privada es igual a 2.04 segundos y, si la imagen procede de la BD pública, el valor de este tiempo es de 1.47 segundos.

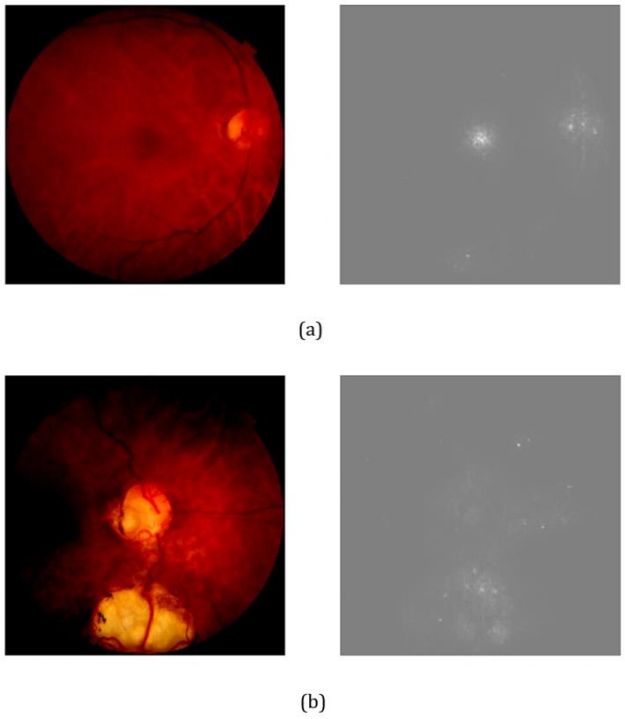


Figura 4.18. Ejemplos de mapas obtenidos con SmoothGrad (con σ=0.10) en la base de datos privada para (a) imagen sin patología y (b) imagen patológica.

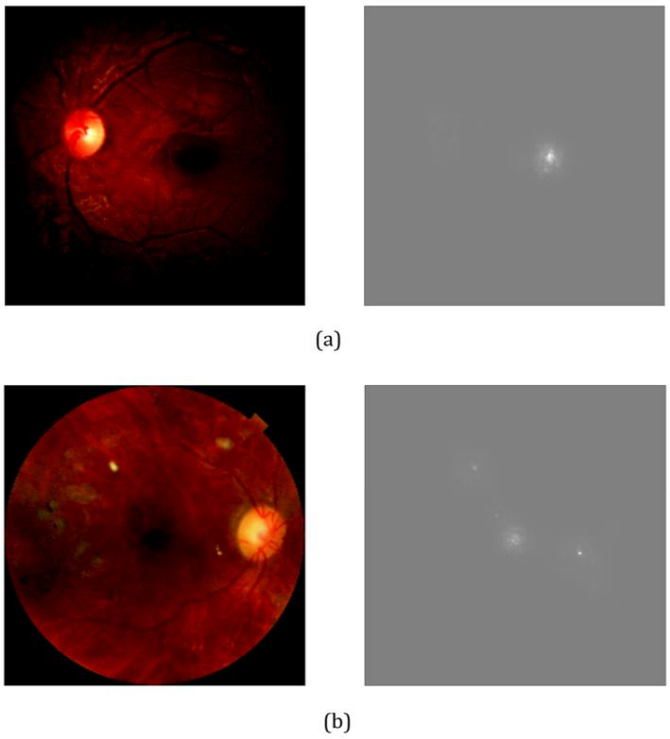


Figura 4.19. Ejemplos de mapas obtenidos con SmoothGrad (con σ=0.10) en la base de datos pública para (a) imagen sin patología y (b) imagen patológica.

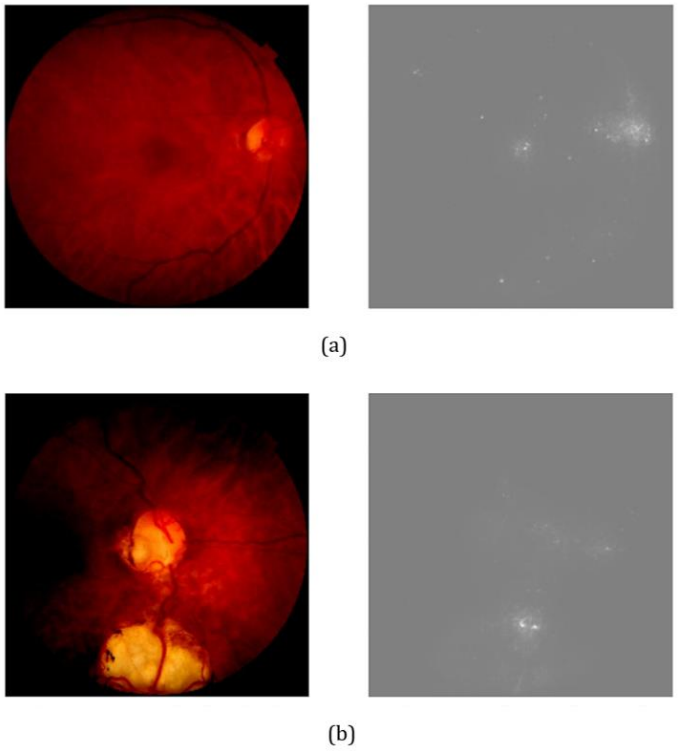


Figura 4.20. Ejemplos de mapas obtenidos con SmoothGrad (con σ=0.15) en la base de datos privada para (a) imagen sin patología y (b) imagen patológica.

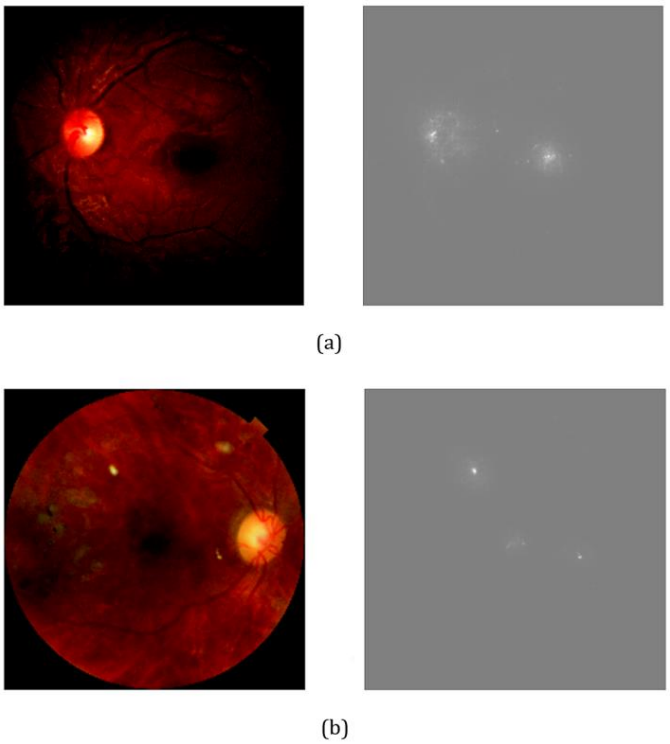


Figura 4.21. Ejemplos de mapas obtenidos con SmoothGrad (con σ=0.15) en la base de datos pública para (a) imagen sin patología y (b) imagen patológica.

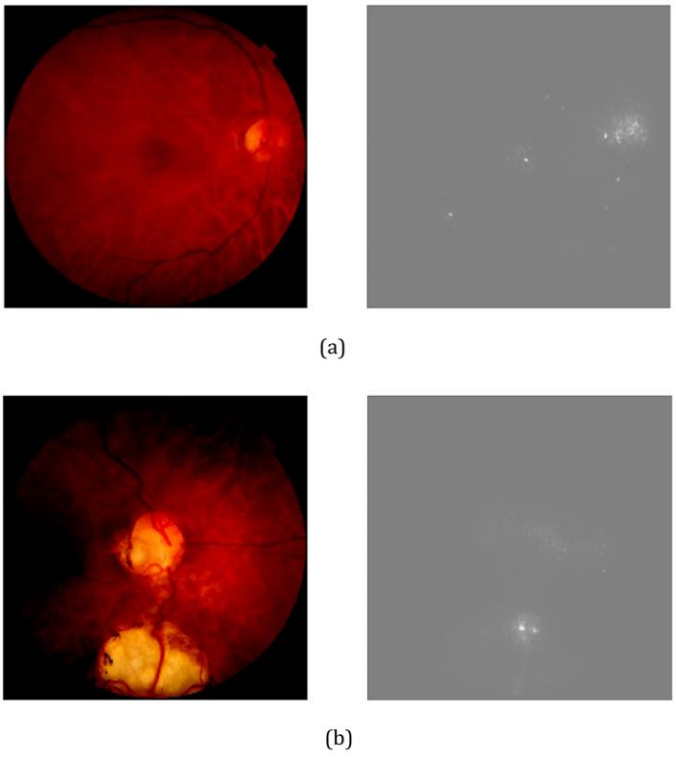


Figura 4.22. Ejemplos de mapas obtenidos con SmoothGrad (con σ=0.20) en la base de datos privada para (a) imagen sin patología y (b) imagen patológica.

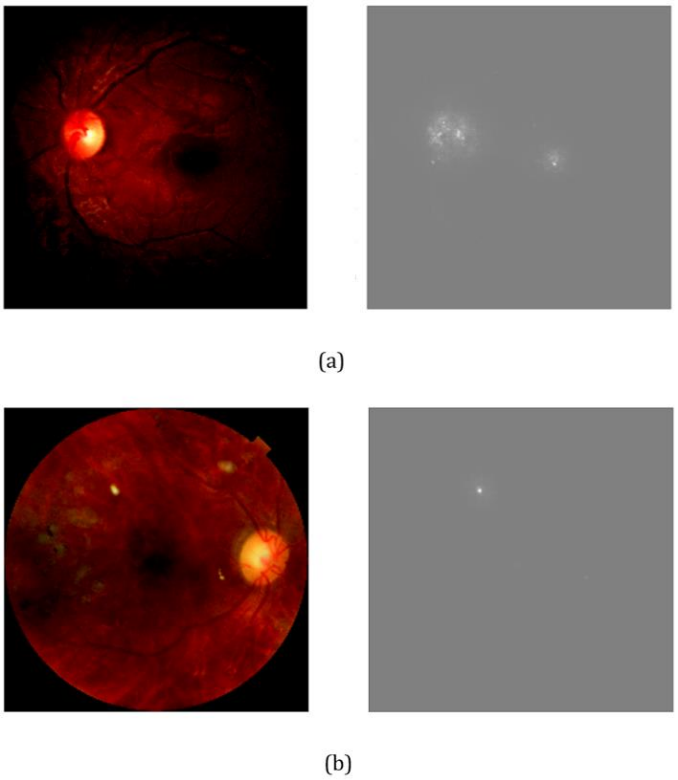
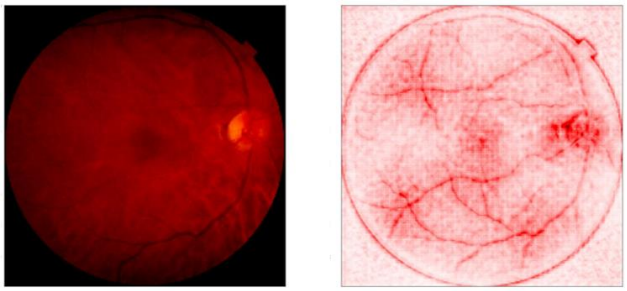


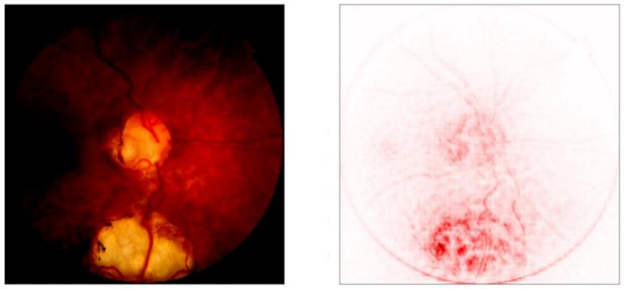
Figura 4.23. Ejemplos de mapas obtenidos con SmoothGrad (con σ=0.20) en la base de datos pública para (a) imagen sin patología y (b) imagen patológica.

	Nivel de ruido (σ)	**Tiempo medio de análisis de una imagen (segundos)**
BD privada	10%	110.99
	15%	78.22
	20%	96.67
BD pública	10%	89.36
	15%	77.23
	20%	93.21

Tabla 4.4. Tiempo medio de análisis de una imagen para diferentes valores de σ (nivel de ruido).

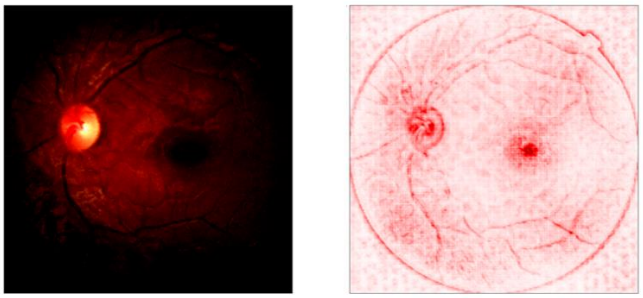


(a)

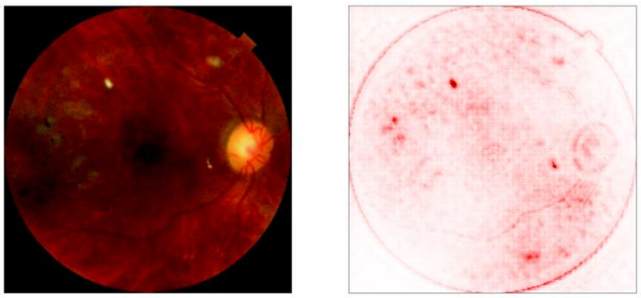


(b)

Figura 4.24. Ejemplos de mapas obtenidos con DeepTaylor (bounded) en la base de datos privada para (a) imagen sin patología y (b) imagen patológica.

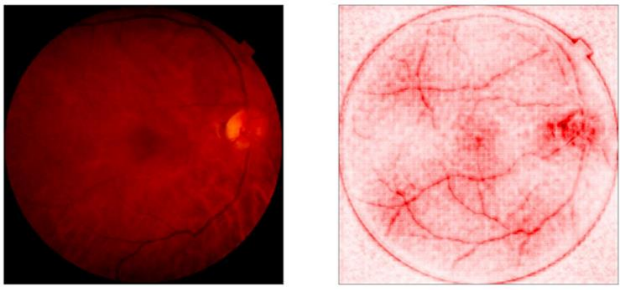


(a)

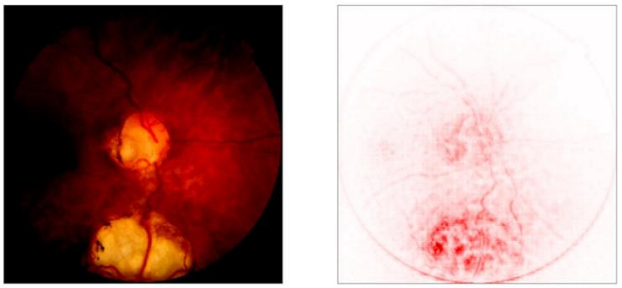


(b)

Figura 4.25. Ejemplos de mapas obtenidos con DeepTaylor (bounded) en la base de datos pública para (a) imagen sin patología y (b) imagen patológica.

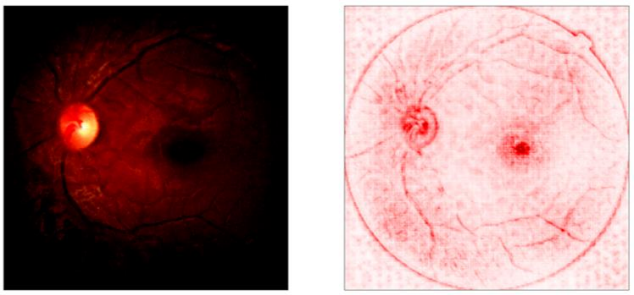


(a)

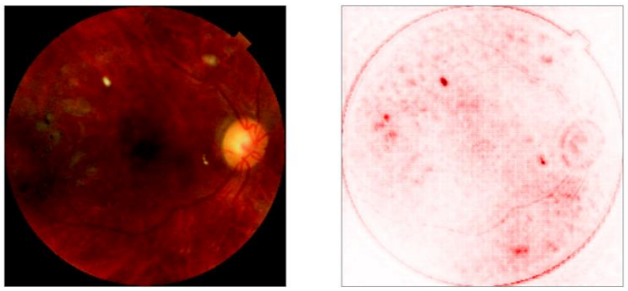


(b)

Figura 4.26. Ejemplos de mapas obtenidos con DeepTaylor (unbounded) en la base de datos privada para (a) imagen sin patología y (b) imagen patológica.



(a)



(b)

Figura 4.27. Ejemplos de mapas obtenidos con DeepTaylor (unbounded) en la base de datos pública para (a) imagen sin patología y (b) imagen patológica.

4.3.2.6. Resultados con LRP-ε

La regla LRP-ε se diferencia del resto por la existencia de un pequeño termino positivo ϵ en el denominador de la expresión para calcular las puntuaciones de relevancia. Dicho término garantiza que solo los factores explicativos más destacados conseguían valores de relevancia elevados, lo que contribuía todo ello a obtener explicaciones menos ruidosas y más dispersas en términos de las características de entrada. Concretamente, en este trabajo, se han obtenido los mapas de atribución correspondientes a valores de iguales a 1e-07, 0.01 y 1. Estos valores se han elegido puesto que ya han sido utilizados en otros estudios previos de clasificación de imágenes (Alber et al., 2019; Seçkin Ayhan et al., 2021; Weith et al., 2018).

En la Figura 4.28 y en la Figura 4.29, se muestran los mapas de atribución correspondientes a un valor de $\varepsilon = 1 \cdot 10^{-7}$ para el caso de la BD privada y de la BD pública, respectivamente. De igual manera, los mapas correspondientes a $\varepsilon = 0.01$, se muestran en la Figura 4.30 y en la Figura 4.31 y, por último, aquellos asociados a un valor de ε = 1, son los que se observan en la Figura 4.32 y en la Figura 4.33. Finalmente, los tiempos medios de análisis de una imagen para cada caso implementado, se especifican en la Tabla 4.5.

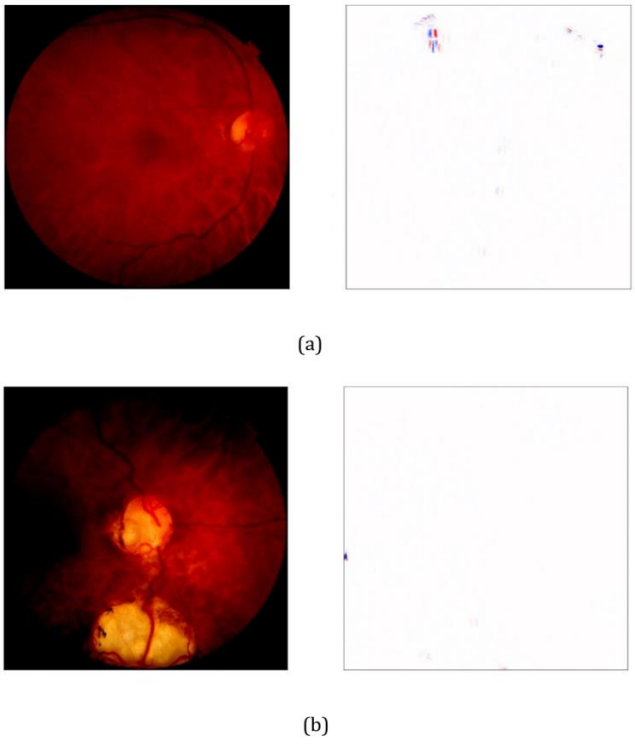


Figura 4.28. Ejemplos de mapas obtenidos con LRP- ε (con ε=1e-07) en la base de datos privada para (a) imagen sin patología y (b) imagen patológica.

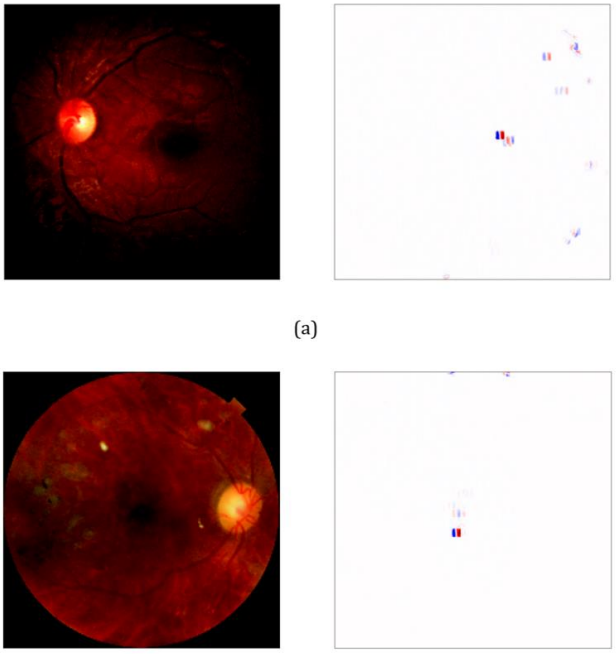


Figura 4.29. Ejemplos de mapas obtenidos con LRP- ε (con ε=1e-07) en la base de datos pública para (a) imagen sin patología y (b) imagen patológica.

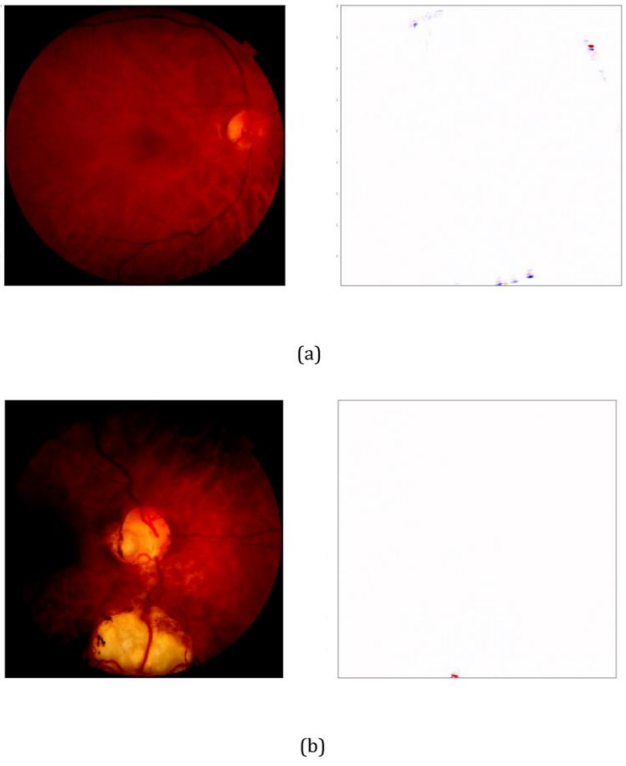


Figura 4.30. Ejemplos de mapas obtenidos con LRP- ε (con ε=0.01) en la base de datos privada para (a) imagen sin patología y (b) imagen patológica.

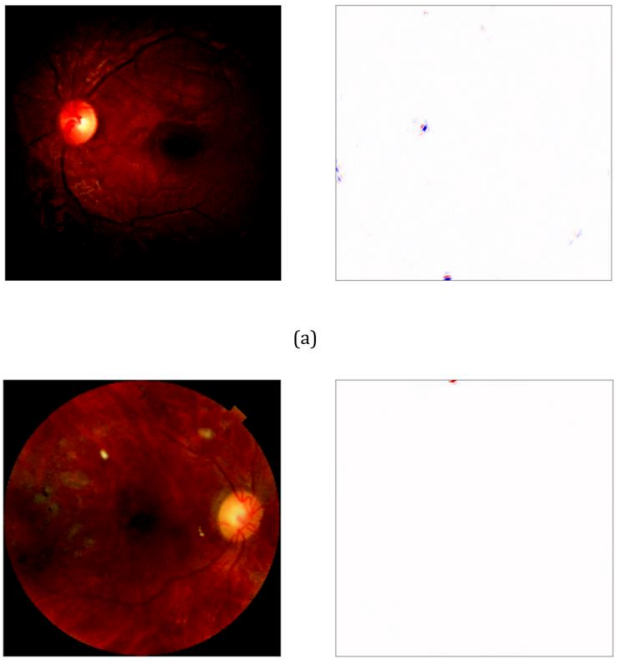


Figura 4.31. Ejemplos de mapas obtenidos con LRP- ε (con ε=0.01) en la base de datos pública para (a) imagen sin patología y (b) imagen patológica.

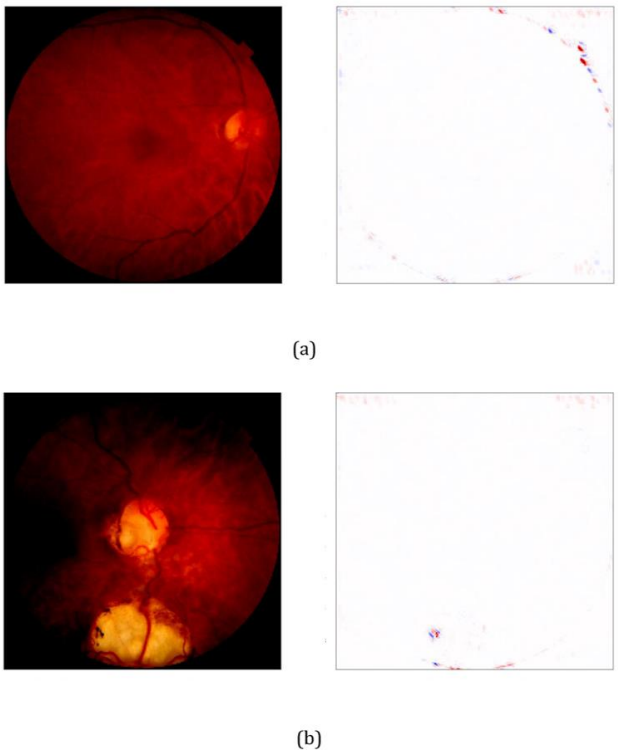


Figura 4.32. Ejemplos de mapas obtenidos con LRP- ε (con ε=1) en la base de datos privada para (a) imagen sin patología y (b) imagen patológica.

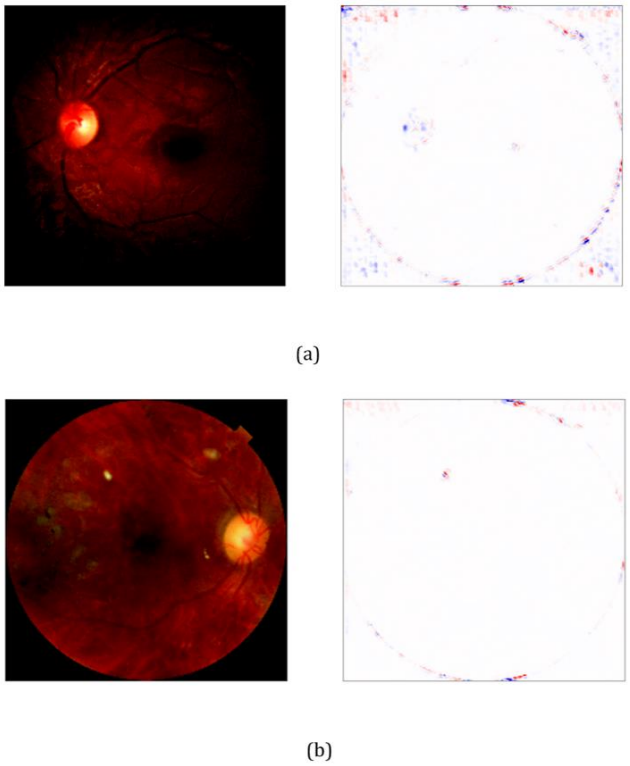


Figura 4.33. Ejemplos de mapas obtenidos con LRP- ε (con ε=1) en la base de datos pública para (a) imagen sin patología y (b) imagen patológica.

	Epsilon (ε)	**Tiempo medio de análisis de una imagen (segundos)**
BD privada	1e-07	2.23
	0.01	2.42
	1	2.80
BD pública	1e-07	1.22
	0.01	1.45
	1	2.00

Tabla 4.5. Tiempo medio de análisis de una imagen para diferentes valores de épsilon.

Capítulo 5

Discusión

5.1. Introducción

A lo largo de este capítulo, se llevará a cabo la discusión acerca de los resultados que se han obtenido en este trabajo. Primero, se realizará una comparación de los resultados con las diferentes bases de datos empleadas, para así poder seleccionar con cuál se obtienen mejores métricas de evaluación. Seguidamente, se analizarán los mapas obtenidos con cada método XAI implementado, con el objetivo de comprobar cuál es el que mejor explica las predicciones de la red. Posteriormente, se discutirán dichos mapas para cada base de datos concreta. Finalmente, este capítulo se culminará realizando una comparación de resultados entre los estudios que ya existen en la literatura actual relacionados con la tarea específica de este TFM y el método que se utiliza a lo largo del mismo.

5.2. Detección de la presencia de patología

Para la realización de este TFM, se ha empleado DenseNet-121 como arquitectura base de la CNN, puesto que fue la que mejores resultados arrojó en nuestro trabajo previo, en el cual se realizó una comparación entre diferentes arquitecturas para detección automática de patología en retinografías. En la fase de entrenamiento se han configurado los parámetros del modelo. Los hiperparámetros se ajustaron de manera empírica teniendo en consideración los valores utilizados en otros estudios previos. La evaluación final del modelo se ha llevado a cabo mediante dos conjuntos de test diferentes, uno de ellos correspondiente a una BD privada (proporcionada por el GIB de la UVa) y el otro correspondiente a una BD pública (RFMiD). Gracias a estos dos grupos de datos independientes se ha podido comprobar la capacidad que tiene la red para generalizar. Los resultados obtenidos aparecen especificados en la Tabla 4.2 del capítulo anterior.

En ambos casos, se puede comprobar que los resultados obtenidos han alcanzado un alto rendimiento. No obstante, se consiguen mejores resultados en las imágenes procedentes de la BD privada respecto a las de la BD pública, puesto que todas las métricas alcanzan un valor más elevado. En el primer caso, se obtiene una precisión del 99%, una sensibilidad igual al 100% (lo que significa que no existe ningún falso negativo, es decir, en ningún caso el sistema detecta que una imagen es sana cuando verdaderamente contiene algún signo patológico) y una especificidad del 98%. Por lo tanto, en este caso todas las retinografías patológicas son detectadas de manera correcta. Además, dado que la precisión da cuenta del número de aciertos con respecto al total de datos, un valor del 99% significa que el sistema solo ha fallado en la detección de una imagen de fondo de ojo (puesto que el conjunto de test de esta BD contiene 100 imágenes). La razón de ello puede deberse a que, en el entrenamiento, la red haya aprendido a detectar como patológica una imagen muy parecida a la que aparece etiquetada como normal en el conjunto de test.

En la Figura 5.1 se muestra la imagen detectada erróneamente junto con la posible patológica con la que se podría haber confundido como consecuencia de su gran similitud. Para obtener más información sobre el rendimiento de la clasificación, se ha analizado la curva ROC, que, tal y como se ha comentado, da cuenta de la relación entre la sensibilidad (tasa de verdaderos positivos) y la especificidad (tasa de falsos positivos). También, se ha calculado el AUC puesto que es una métrica muy utilizada en los problemas de cribado de patologías en el campo sanitario. Para este conjunto de test, se ha obtenido un AUC igual a 1, por lo que se comprueba que el sistema desarrollado ofrece un alto rendimiento.

En el caso de las imágenes procedentes de la BD pública, los resultados son algo inferiores, pues se ha alcanzado una precisión igual al 90.93%, una sensibilidad igual

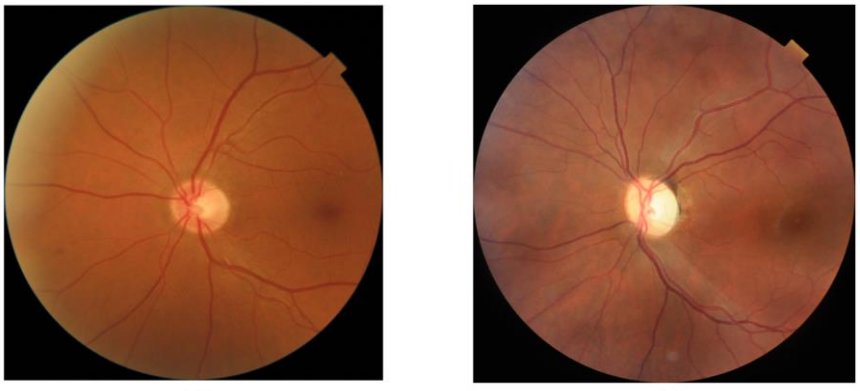


Figura 5.1. (a) Imagen normal del conjunto de test correspondiente a la BD privada y (b) Imagen patológica del conjunto de entrenamiento.

al 91.30% y una especificidad del 89.55%(ver Tabla 4.2). Dado que este conjunto de datos ahora está formado por 640 imágenes, estos valores indican que la red solo ha detectado bien 120 imágenes (de 134) correspondientes a sujetos sanos y 462 imágenes (de 506) correspondientes a sujetos con signos patológicos. Dicho de otro modo, el modelo ha fallado en la detección de 14 retinografías sanas y 44 retinografías patológicas. Luego, tal y como se observa en la matriz de confusión correspondiente (Figura 4.5b), alrededor del 90% de imágenes de fondo de ojo normales y alrededor del 91% de patológicas han sido clasificadas correctamente. El hecho de que se hayan conseguido resultados algo mejores en la detección de imágenes patológicas puede deberse al desbalanceo de clases, pues las clases con mayor número de aciertos se corresponden con aquellas que contienen un mayor número de imágenes en la BD. Además, el hecho de que las imágenes de la BD RFMiD hayan sido etiquetadas por diferentes oftalmólogos, hace que sea posible la discordancia en la clasificación y que, por lo tanto, la misma imagen pueda ser clasificada de manera diferente. En este sentido, el hecho de que la arquitectura haya fallado en la predicción de imágenes sanas no es casualidad, puesto que la red puede haber aprendido a detectar, como imagen patológica, otra muy parecida clasificada como normal en el conjunto de test. En la Figura 5.2, se muestran dos ejemplos de retinografías sanas junto con las posibles imágenes pertenecientes a sujetos patológicos con las que se podría haber confundido dada su gran similitud.

Relacionado con los valores de sensibilidad y especificidad, se calculó también la curva ROC y el valor del AUC para evaluar el sistema desarrollado. En este caso se obtuvo un AUC algo inferior, igual al 0.97. No obstante, también es un valor bastante elevado que demuestra que el sistema tiene una gran capacidad de detección de patologías.

5.3. Comparación entre diferentes métodos XAI

En los entornos clínicos reales, resulta imprescindible entender por qué los modelos de DL toman unas decisiones u otras (Van der Velden et al., 2022). Aunque los resultados obtenidos sobre los conjuntos de tests sean positivos, este criterio resulta insuficiente para el uso de los modelos en aplicaciones médicas reales. De ahí la gran importancia que tiene XAI en este contexto. Gracias a las técnicas que engloba, es posible interpretar las predicciones realizadas por los modelos.

En este trabajo se han implementado seis métodos de atribución diferentes, los cuales tienen por objetivo determinar la contribución de una característica de entrada en la neurona de salida de la clase correcta. Estas contribuciones o

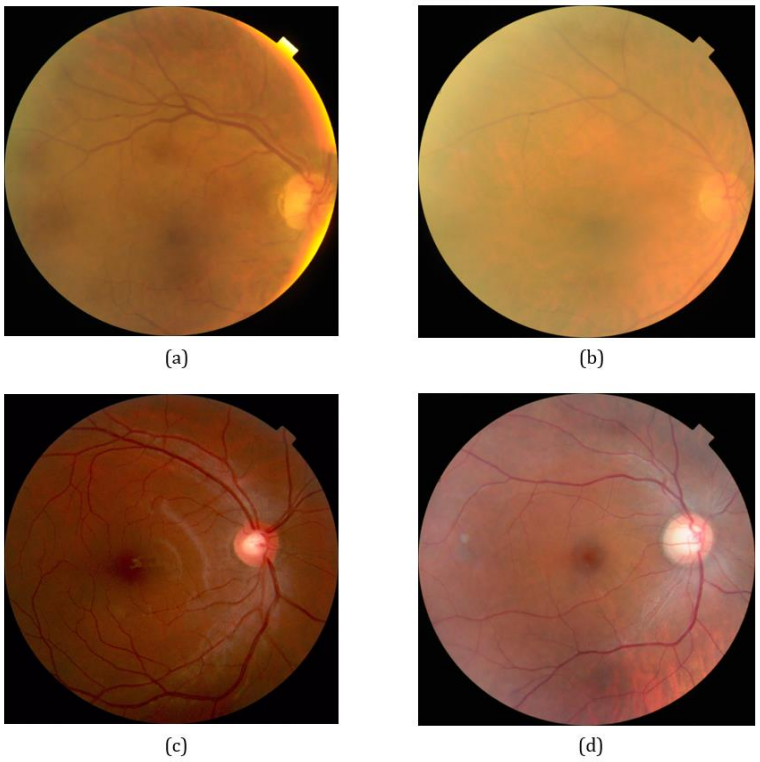


Figura 5.2. (a) Primer ejemplo de imagen normal del conjunto de test correspondiente a la BD pública, (b) Imagen patológica del conjunto de entrenamiento similar al primer ejemplo, (c) Segundo ejemplo de imagen normal del conjunto de test correspondiente a la BD pública y (d) Imagen patológica del conjunto de entrenamiento similar al segundo ejemplo.

atribuciones se disponen en mapas de atribución o mapas de calor. En el capítulo anterior, se han mostrado ejemplos de mapas obtenidos con cada método XAI empleado. Para realizar una mejor comparación entre ellos, se van a utilizar aquellos correspondientes a las mismas retinografías originales.

En las visualizaciones obtenidas con SHAP, se muestra, de izquierda a derecha, la imagen de fondo de ojo original analizada y los mapas SHAP obtenidos para cada clase, siendo la de la izquierda la correspondiente a la clase negativa o ausencia de patología y la de la derecha la de la clase positiva o presencia de patología. Además, en estos mapas, el color rojo indica que la característica tiene una contribución positiva para la activación de la neurona de salida, mientras que el color azul significa una contribución negativa. Esto significa que los píxeles que aparecen en rojo representan aquellos que pertenecen a una clase determinada y los azules lo descartan. En la Figura 5.3 se muestra un ejemplo de visualización SHAP para una retinografía detectada correctamente y otra con la que el modelo ha fallado. Concretamente, el primer caso se trata de una imagen sana y, en el segundo caso, se detecta como retinografía normal cuando verdaderamente es patológica.

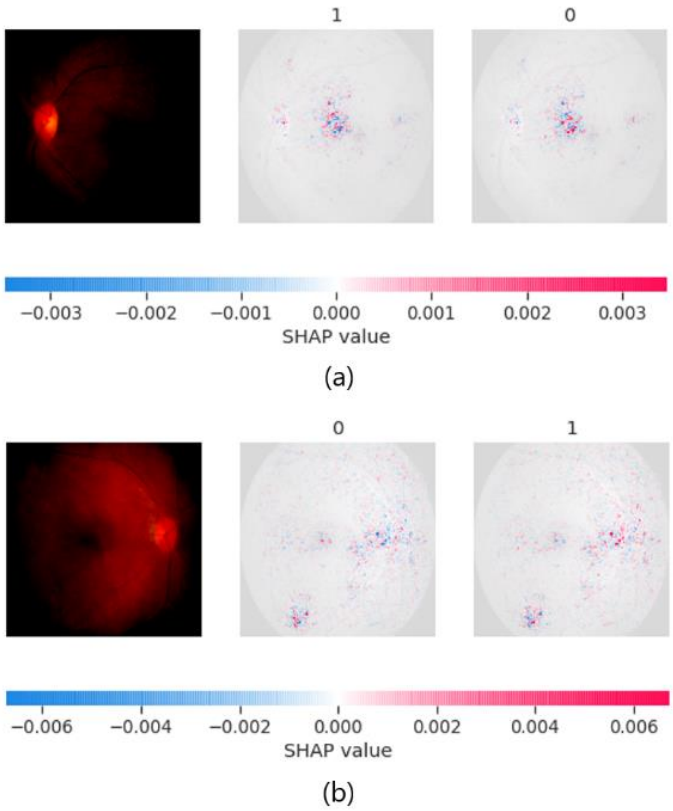


Figura 5.3. Visualización SHAP de (a) imagen normal detectada correctamente y (b) imagen normal detectada de manera errónea.

En la Figura 5.3a, se comprueba que la clase correspondiente a la ausencia de patología es la que mayor porcentaje de píxeles rojos presenta. Además, estos píxeles aparecen en regiones que no contienen lesiones aparentes. Sin embargo, esto no ocurre en la Figura 5.3b, donde ahora la clase asociada a presencia de enfermedad es la que más píxeles rojos presenta y aparecen varias regiones coloreadas, pudiendo indicar la presencia de alguna lesión tales como exudados o hemorragias. Eso puede deberse a que la red ha detectado erróneamente algún cambio de textura en la imagen. De ahí que el sistema la haya detectado como imagen patológica cuando verdaderamente no lo es.

El método *Input x Gradient*, como se ha comentado, obtiene cada valor de atribución calculando el gradiente de la salida multiplicado por la entrada, puesto que se basa en la idea de que el primero nos da cuenta de la importancia de una dimensión y, la entrada nos permite saber la intensidad con la que se expresa dicha dimensión en la imagen. En la Figura 5.4 se muestran los mapas de atribución obtenidos con este método para los mismos casos que con SHAP. Tanto en el primero (Figura 5.4a) como en el segundo (Figura 5.4b), se observa que existen más píxeles rojos que azules, luego hay más contribuciones positivas que negativas para la activación de

la neurona de salida determinada. Además, en el primer mapa se observa que solo aparece una región coloreada en el centro de la retinografía, que se corresponde con la parte conocida como fóvea. Sin embargo, en el segundo caso aparecen coloreadas varias regiones con una mayor densidad de píxeles rojos. En este último caso, además de aparecer coloreada la zona de la mácula, también aparecen varios píxeles rodeando la zona del DO y en la región inferior izquierda, pudiéndose tratar de lesiones tales como MAs, daños en el nervio óptico (propios del glaucoma) o EXs, respectivamente. Por ello, en el primer caso la imagen ha sido clasificada como sana mientras que en el segundo, se ha detectado como patológica. No obstante, esta técnica ofrece mapas de atribución que presentan poca nitidez (aparece cierto ruido) como consecuencia de que el gradiente proporciona información local cuando los cambios en la entrada son pequeños.

En cuanto al método IG, éste obtiene cada valor de atribución mediante el cálculo de una integral, que en la práctica se suele aproximar mediante *Riemman*. Esta aproximación implica la existencia de un hiperparámetro que se puede variar, que hace referencia al número de pasos que se desean emplear para calcular los gradientes. Tal y como ya se comentó, en nuestro caso, se realizaron pruebas con

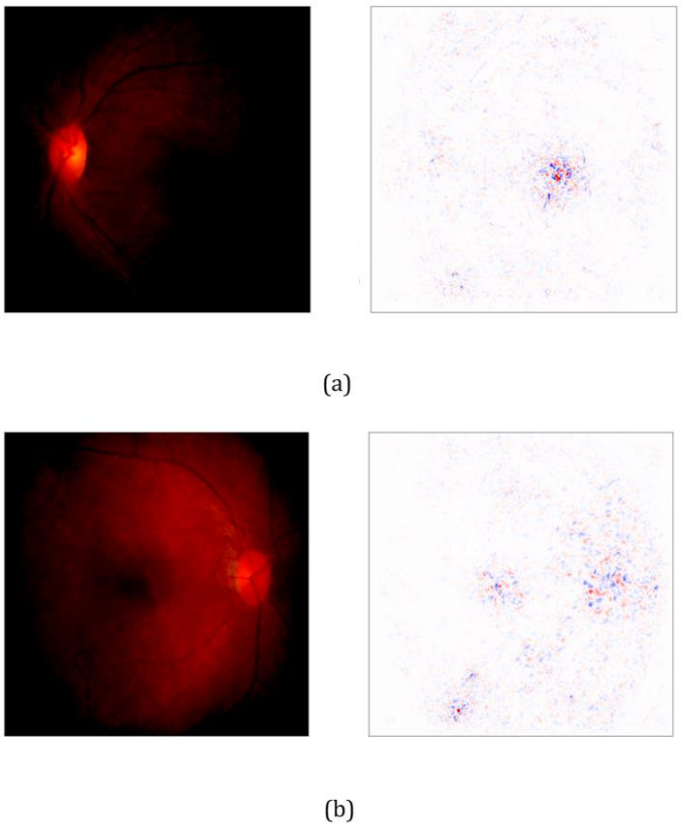


Figura 5.4. Visualización Input x Gradient de (a) imagen normal detectada correctamente y (b) imagen normal detectada de manera errónea.

tres valores diferentes: 64 (Figura 5.5), 100 (Figura 5.6) y 160 (Figura 5.7). Como se puede observar, los mapas de atribución que presentan menos ruido son los obtenidos en el último caso, puesto que, al calcular un mayor número de gradientes para obtener cada valor de atribución, los resultados que se consiguen son más precisos. De igual manera que con los dos métodos anteriores, en los mapas de la izquierda aparecen menos regiones coloreadas respecto a los mapas presentados a la derecha. Por esta razón, en el primer caso se detecta como una imagen asociada a un sujeto sano, mientras que, en el segundo, la imagen se clasifica como perteneciente a un sujeto con patología. No obstante, esto último se debe a que la red ha aprendido en el entrenamiento características de una imagen patológica muy similar a dicha retinografía normal, de ahí que la red se haya equivocado.

Respecto a *DeepTayor*, se trata de uno de los métodos de atribución que solo proporciona pruebas positivas (aparecen únicamente píxeles rojos en los mapas). Para este caso, se ha probado a limitar los valores de la señal de entrada (en el rango [-1, 1]), caso conocido como *bounded* y a no limitarlos, caso denominado *unbounded*. Los mapas obtenidos se muestran en la Figura 5.8 y en la Figura 5.9, respectivamente. Tal y como se puede comprobar, apenas existen diferencias entre

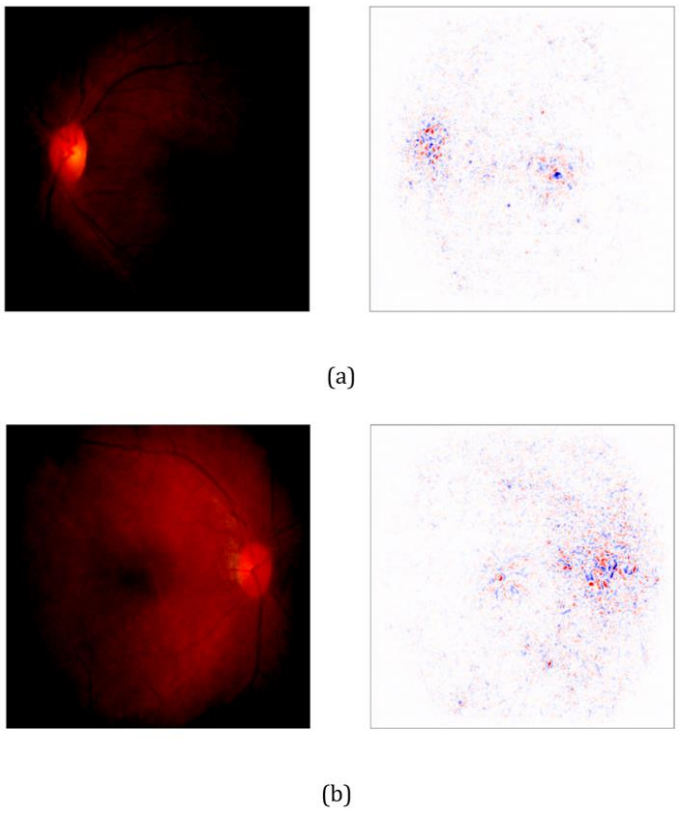


Figura 5.5. Visualización Integrated Gradients (con m=64) de (a) imagen normal detectada correctamente y (b) imagen normal detectada de manera errónea.

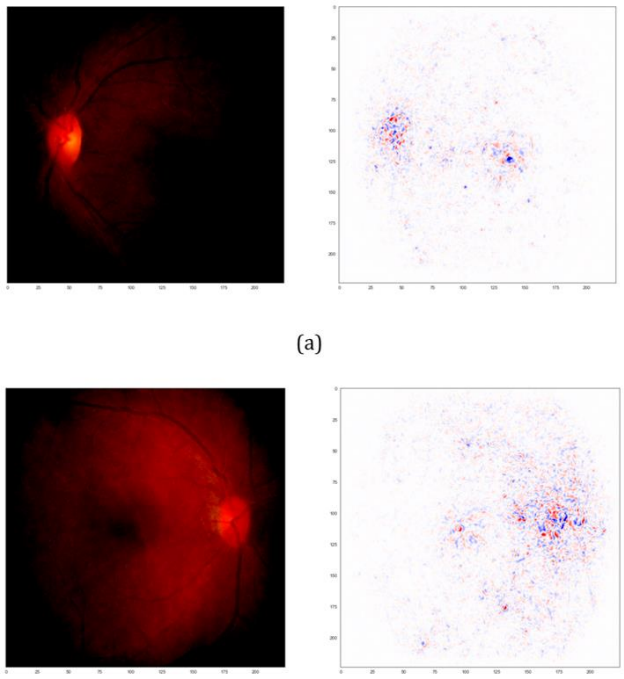


Figura 5.6. Visualización Integrated Gradients (con m=100) de (a) imagen normal detectada correctamente y (b) imagen normal detectada de manera errónea.

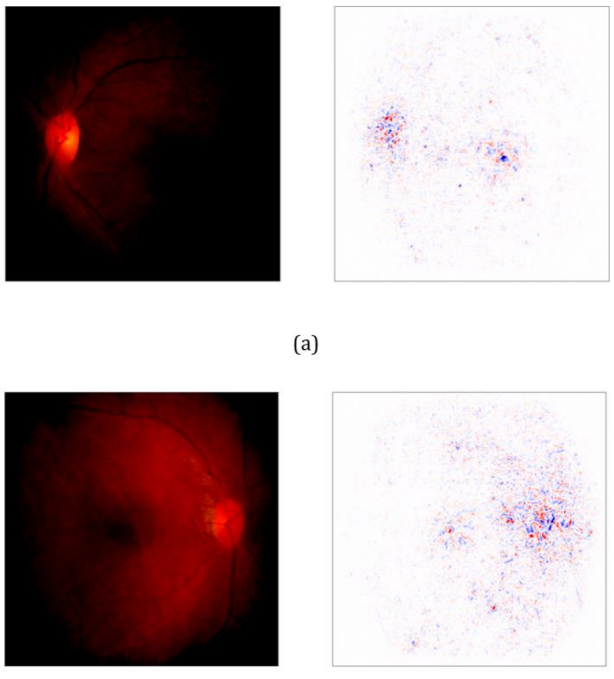


Figura 5.7. Visualización Integrated Gradients (con m=160) de (a) imagen normal detectada correctamente y (b) imagen normal detectada de manera errónea.

ambos casos puesto que se obtienen mapas prácticamente idénticos. Además, se observan pocas diferencias respecto al caso que se detecta correctamente (mapas de la izquierda) y al caso clasificado de manera errónea (mapas presentados a la derecha). En ambos es posible identificar la región del DO, la fóvea y los vasos sanguíneos. Además, en la Figura 5.8b y en la Figura 5.9b, se observa que la región del DO está coloreada ligeramente con mayor intensidad alrededor de la primera región (el DO). Esta podría ser la razón por la que la red la ha detectado como imagen que presenta patología. No obstante, este método XAI no ofrece resultados suficientemente claros que justifiquen las predicciones de la red, en comparación con los anteriores, con los que se pueden observar varias regiones coloreadas, que la red ha podido detectar como lesiones características de diferentes enfermedades.

SmoothGrad trata de minimizar el problema de que el gradiente puede fluctuar bruscamente a pequeñas escalas. Para ello, muestrea varias imágenes similares añadiendo ruido a la imagen original y finalmente, calcula la media de los mapas de sensibilidad resultantes de cada imagen muestreada. Cuenta con dos hiperparámetros que son el número de muestras (n) y el nivel de ruido añadido (σ). En este trabajo se ha fijado n=50 y se han obtenido los mapas de atribución correspondientes a σ=0.1 (Figura 5.10), σ=0.15 (Figura 5.11) y σ=0.2 (Figura 5.12).

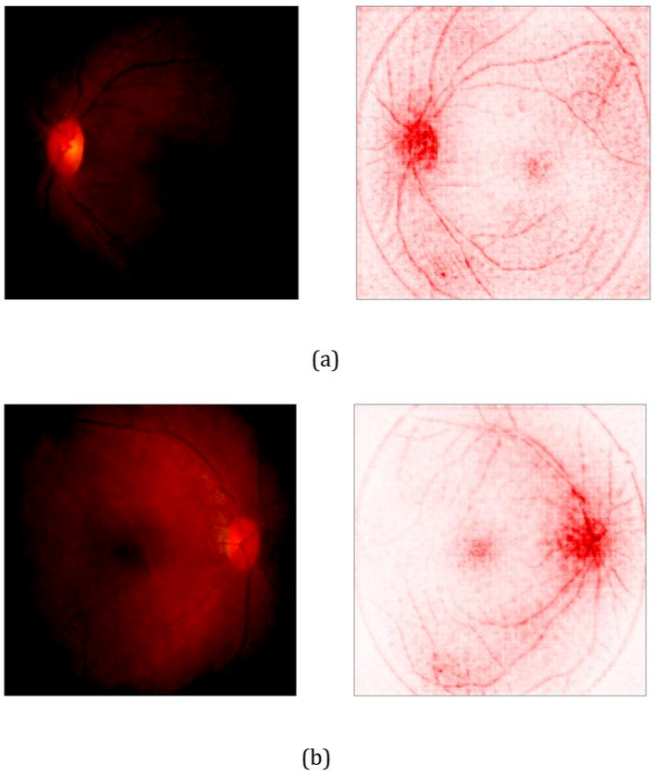


Figura 5.8. Visualización DeepTaylor (bounded) de (a) imagen normal detectada correctamente y (b) imagen normal detectada de manera errónea.

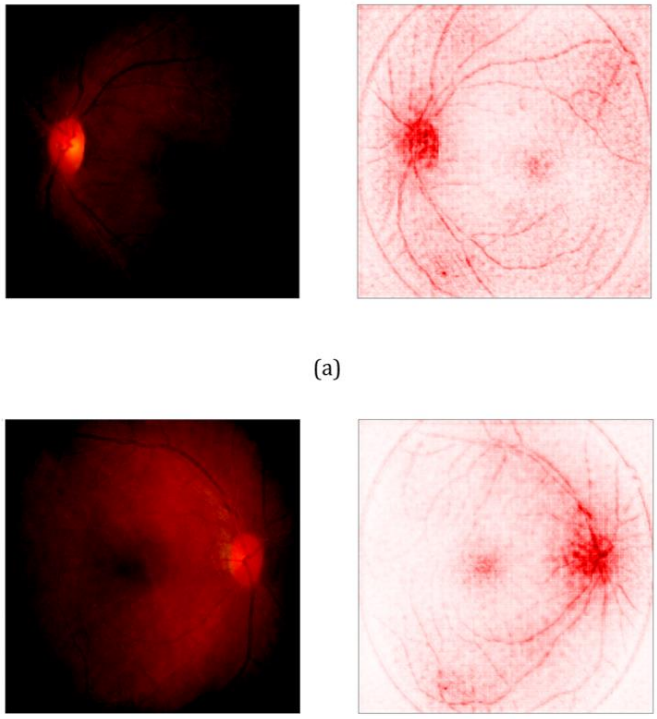
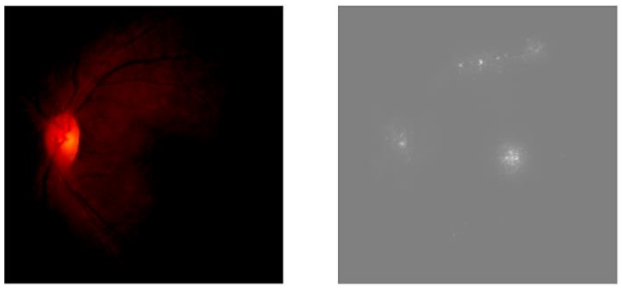


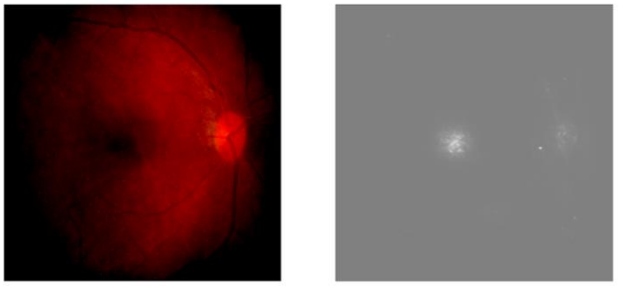
Figura 5.9. Visualización DeepTaylor (unbounded) de (a) imagen normal detectada correctamente y (b) imagen normal detectada de manera errónea.

Tal y como se puede observar, con este método no aparecen píxeles rojos ni azules, si no que solo las contribuciones más significativas se colorean en blanco. Teniendo esto en cuenta, los resultados más explicativos se obtienen en el segundo caso, puesto que la adición de un 10% de ruido provoca que aparezcan coloreadas regiones que no son muy significativas. Lo contrario sucede con la adición de un nivel de ruido del 20%, es decir, solo las regiones muy significativas son las que aparecen en blanco. Un valor intermedio (igual al 15%) permite observar cómo en el mapa de la izquierda solo aparece como contribución relevante la zona del DO mientras que, en el de la derecha, se pueden observar varias regiones en blanco que la red ha interpretado como posibles lesiones. Por ello, en este último caso, la retinografía ha sido clasificada como patológica.

Finalmente, el método LRP calcula puntuaciones de relevancia que propaga de manera inversa por la red utilizando reglas heurísticas de propagación que se aplican a cada capa. En nuestro caso, se ha empleado la regla épsilon, que añade un término positivo ϵ (en el denominador de la expresión para calcular las relevancias). Su papel es absorber cierta relevancia cuando las contribuciones a la activación de

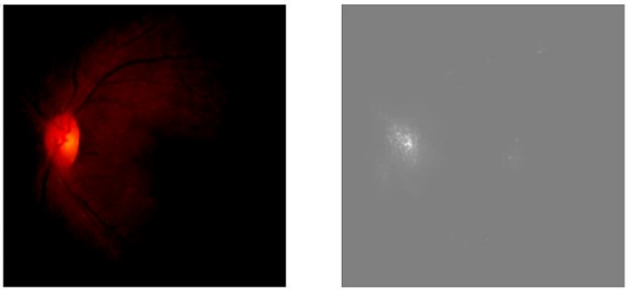


(a)

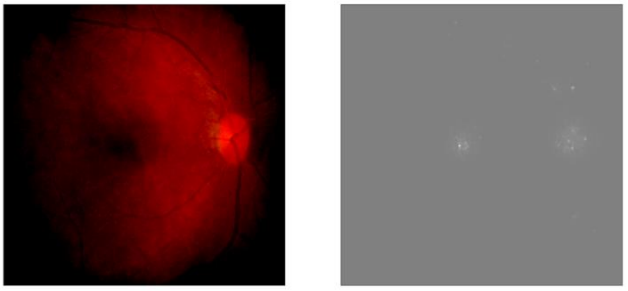


(b)

Figura 5.10. Visualización Smoothgrad (con σ=0.10) de (a) imagen normal detectada correctamente y (b) imagen normal detectada de manera errónea.



(a)



(b)

Figura 5.11. Visualización Smoothgrad (con σ=0.15) de (a) imagen normal detectada correctamente y (b) imagen normal detectada de manera errónea.

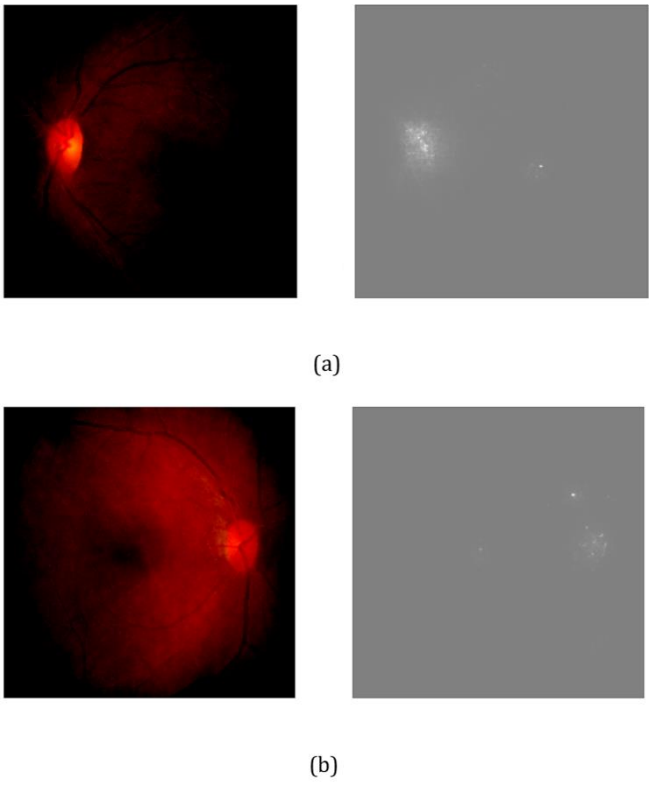


Figura 5.12. Visualización Smoothgrad (con σ=0.20) de (a) imagen normal detectada correctamente y (b) imagen normal detectada de manera errónea.

la neurona de salida son débiles o contradictorias. Por tanto, a medida que aumenta, solo los factores explicativos más destacados sobreviven a la absorción. En nuestro caso se han hecho pruebas con tres valores diferentes de ε: 10^{-7}(Figura 5.13), 0.01 (Figura 5.14) y 1 (Figura 5.15). No obstante, tal y como se puede comprobar, con ninguno de ellos se obtienen resultados claros. Aunque es cierto que, con valores pequeños de ϵ, aparecen más regiones coloreadas respecto al caso con mayor valor (en el que se detecta, de manera incorrecta, el fondo de la imagen como relevante), en ninguno de los mapas de atribución obtenidos, se muestra de manera clara por qué la red detecta la retinografía de la izquierda como sana y la de la derecha como patológica. Por lo tanto, esta última técnica XAI no permite obtener resultados relevantes para la tarea bajo estudio. En comparación con el resto, los mapas de atribución que se consiguen con ella, son los que ofrecen menor información sobre las predicciones realizadas por la red.

Por último, en cuanto al tiempo medio que se tarda en analizar cada imagen con los diferentes métodos, se observa que, en la mayoría de los casos, estos tiempos son algo inferiores en las imágenes de test procedentes de la BD pública respecto a la BD privada. Además, en cuanto a los métodos XAI, el que menos tiempo tarda es

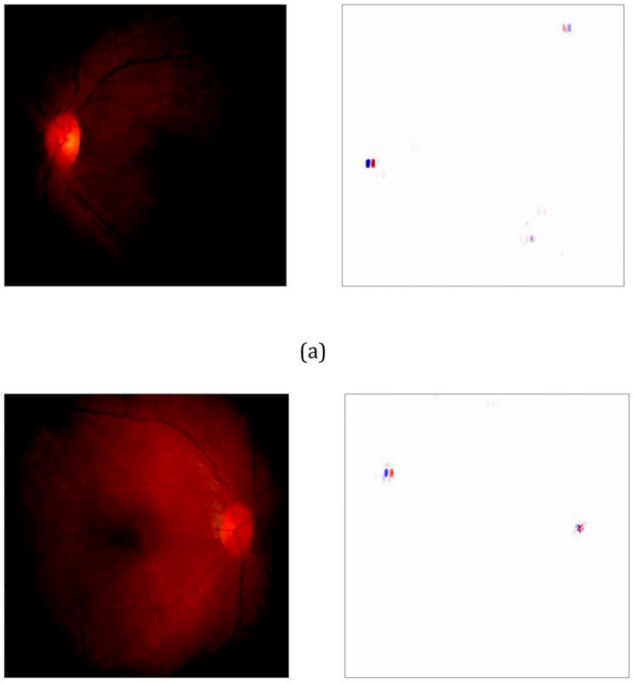


Figura 5.13. Visualización LRP-epsilon (con ε=1e-07) de (a) imagen normal detectada correctamente y (b) imagen normal detectada de manera errónea.

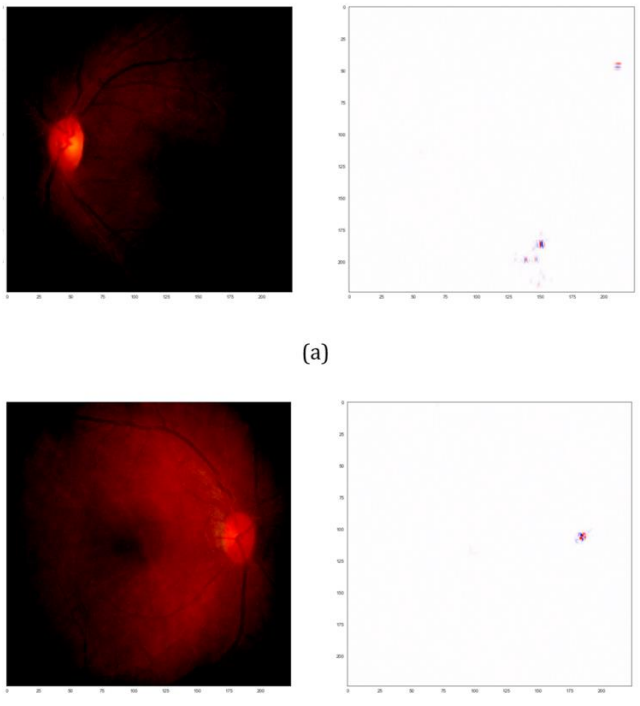


Figura 5.14. Visualización LRP-epsilon (con ε=0.01) de (a) imagen normal detectada correctamente y (b) imagen normal detectada de manera errónea.

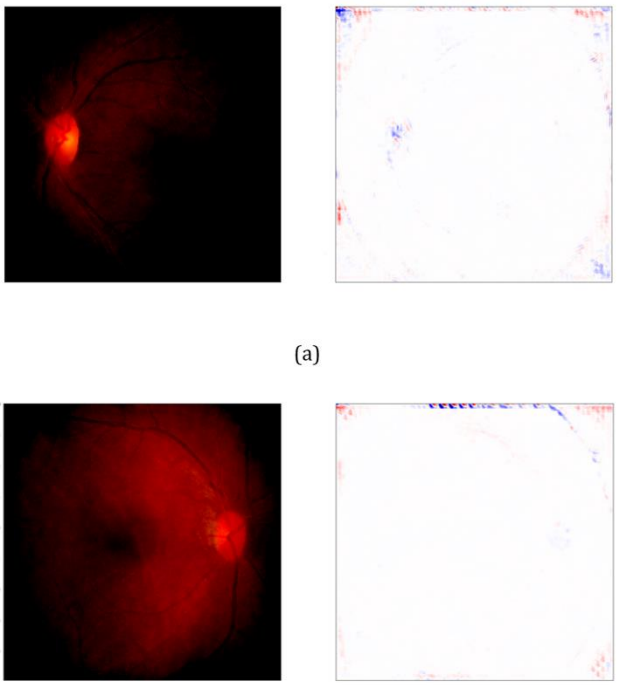


Figura 5.15. Visualización LRP-epsilon (con ε=1) de (a) imagen normal detectada correctamente y (b) imagen normal detectada de manera errónea.

DeepTaylor bounded, puesto que únicamente tarda 1.07 segundos en analizar una imagen de la BD pública y 1.75 segundos cuando analiza una procedente de la BD privada. No obstante, estos tiempos son similares a los que consigue *Input x Gradient* (tarda 1.22 segundos en analizar una retinografía de la BD pública y 1.82 segundos en analizar una de la BD privada). Sin embargo, es IG, con $m = 160$ pasos, el método cuyo tiempo medio de análisis por imagen es claramente superior. Concretamente, este último tarda en analizar cada imagen de la BD pública, 258.91 segundos y 262.16 segundos de la BD privada. Por tanto, las diferencias son muy notables. Esto se debe a que este método calcula cada píxel del mapa mediante una integral, por lo que el coste computacional es mayor comparado con otras técnicas. No obstante, a pesar de ello, se ha comprobado que es el que permite explicar mejor los resultados. Por ello, se considera el método de atribución XAI óptimo para nuestra tarea específica. Dicho método ya ha sido utilizado con éxito en estudios previos relacionados centrados en la detección de la RD (Sayres et al., 2019).

5.4. Comparación con estudios previos

En este último apartado del capítulo, se va a realizar una comparación entre los

resultados alcanzados en este TFM y aquellos que se han obtenido en estudios previos relacionados. Dado que el método desarrollado permite el cribado automático de un gran número de patologías, se han analizado estudios que llevan a cabo esta misma tarea. La comparativa realizada se muestra en la Tabla 5.1, donde se especifican los valores de las métricas más comunes (como son el AUC, la precisión, la sensibilidad y la especificidad) para cada estudio analizado y para el que se ha llevado a cabo a lo largo de este trabajo. Con respecto a los métodos XAI, conviene señalar que no se ha encontrado ningún estudio en la literatura donde se apliquen métodos de atribución para explicar los resultados obtenidos en la clasificación automática de patologías en general. Esto se debe al hecho de que el concepto de XAI es todavía bastante reciente.

Lo primero que se puede comprobar es que la mayoría de los estudios no incluyen todas las métricas de evaluación para medir sus resultados. Aunque es cierto que muchos de ellos incluyen el valor del AUC, en el método propuesto en este TFM se han calculado todas las métricas especificadas en la Tabla 5.1. Adicionalmente, se han incluido también las matrices de confusión, con el objetivo de poder analizar mejor los resultados para cada clase discriminada en el problema. También, se puede observar como casi todos los estudios utilizan BBDD públicas para implementar las redes neuronales. Incluso, algunos de ellos emplean varias BBDD para conformar una única que es la que se aplica a la CNN correspondiente. Esto permite disponer de un mayor número de imágenes de fondo de ojo con signos patológicos y, por lo tanto, que la red sea capaz de detectar una gran variedad de patologías oculares. Es por ello por lo que en los trabajos que se utilizan BD de gran tamaño se obtienen muy buenos resultados, como es en el caso de (Sarki et al., 2022), ya que consigue una especificidad y una sensibilidad óptimas iguales al 100%.

Si analizamos el algoritmo propuesto en este TFM, tanto si se emplea la BD privada como la pública, los resultados obtenidos, en cuanto al valor del AUC y de la precisión, superan los de los demás estudios. Además, con la primera BD los valores de todas las métricas son prácticamente los óptimos. No obstante, dado que no se ha empleado en todos los casos la misma BD, la comparación se debe realizar con cierta precaución. En el estudios de (Choi et al., 2017) se lleva a cabo la detección de 9 patologías diferentes para lo cual, se empleó una arquitectura VGG-19 junto con el aprendizaje de transferencia (utilizando la red preentrenada en ImageNet). Los resultados de esta CNN se utilizaron posteriormente en un clasificador *random forest*, encargado de realizar la clasificación final. En (Quellec et al., 2020) se propone el cribado de un mayor número de patologías empleando una arquitectura Inception-V3 también junto con la técnica de *transfer learning* pero, a diferencia de la anterior, son las últimas capas de la CNN (capas *fully-connected*) las encargadas de realizar la clasificación final.

Por otro lado, en (Luo et al., 2021) no se emplea el aprendizaje de transferencia pero se utiliza la técnica CLAHE para mejorar la calidad de las retinografías de la BD seleccionada. Además, la arquitectura EfficientNet se emplea como CNN para detectar las enfermedades oculares más comunes tales como cataratas, DMAE, glaucoma y RD. Finalmente, los mismos autores presentan dos estudios muy parecidos como son (Sarki et al., 2020) y (Sarki et al., 2022). En ambos se pretende detectar la EOD que, como ya se comentó, agrupa a varias patologías oculares pero, en el primer caso, el problema es de clasificación binaria mientras que en el segundo se lleva a cabo una clasificación multiclase. Además, los modelos de CNN empleados son diferentes ya que en el primer estudio se utiliza la arquitectura VGG-16 mientras que, en el segundo, se diseña una CNN profunda ajustando sus hiperparámetros correspondientes. Los resultados que se obtienen son similares en los dos estudios.

En conclusión, el sistema automático diseñado consigue alcanzar resultados de gran calidad en ambos casos. Esto se debe a que el DL ha permitido superar los resultados de los métodos clásicos (tales como los basados en técnicas de procesado de imagen), consiguiendo en menor tiempo un diagnóstico más fiable. No obstante, en ciertas circunstancias estos diagnósticos necesitan ser explicados puesto que, en el campo de la medicina, las predicciones deben ser fiables antes de ser incorporadas a la práctica clínica. Es por ello por lo que, en este trabajo, se ha decidido complementar los resultados con diferentes técnicas XAI. Además, tal y como se ha podido comprobar, apenas existen estudios en la literatura que lo incluyan. Se ha comprobado que solo existen aquellos relacionados con algunas patologías oculares más frecuentes, tales como la RD o el glaucoma. Sin embargo, no se ha encontrado la implementación de métodos XAI para el diagnóstico de cataratas.

Por lo tanto, en este sentido, el marco propuesto en este trabajo es bastante novedoso, ya que es capaz de analizar retinografías procedentes de una gran variedad de patologías, incluyendo desde las más comunes hasta las más raras. Luego, al implementar un sistema tan general, se podría aplicar en un gran número de centros de salud, ya que los oftalmólogos podrían emplearlo con mayor fiabilidad puesto que está diseñado para detectar un amplio rango de afecciones oculares.

Autor/es (año)	Descripción breve del método	Base/s de datos	Resultados sobre el conjunto de test			
			AUC	Precisión	Sensibilidad	Especificidad
Choi et al. (2017)	Detección de 9 patologías diferentes con VGG-19 y clasificador *random forest*	STARE	0.90	-	80.30%	85.50%
Quellec et al. (2020)	Detección de 37 patologías con Inception-V3, PCA y modelo probabilístico	OPHDIAT	0.94	-	-	-
Sarki et al. (2020)	Detección de enfermedades oculares diabéticas leves con VGG-16	*Messidor*, *Messidor-2* y DRISHTI-GS	-	85.94%	-	-
Luo et al. (2021)	Detección general de patologías utilizando CLAHE y EfficientNet	BD privada	0.85	90.08%	94.71%	76.16%
Sarki et al. (2022)	Clasificación multiclase de la EOD con red neuronal profunda	*Messidor*, *Messidor-2*, DRISHTI-GS y *Kaggle cataract*	-	81.33%	-	-
Método propuesto	Detección general de patologías con DenseNet-121 junto con *data augmentation*, *transfer learning*, *fine tuning* y *dropout*	BD privada	0.99	99.00%	100%	98.00%
		RFMiD	0.97	90.93%	91.30%	89.55%

Tabla 5.1.Comparación de resultados en el conjunto de test de los métodos previos relacionados y los correspondientes al método propuesto en este TFM.

Capítulo 6

Conclusiones y líneas futuras

6.1. Introducción

El notable éxito del DL ha despertado el interés de su aplicación al diagnóstico de imágenes médicas. Aunque es cierto que los modelos de DL más avanzados han logrado alcanzar precisiones de nivel humano en la clasificación de diferentes tipos de datos médicos, en la práctica, estos modelos apenas se adoptan debido principalmente a su falta de interpretabilidad. Su carácter de "caja negra" ha planteado la necesidad de desarrollar estrategias para explicar el proceso de decisiones de estos modelos, lo que ha llevado a la creación de XAI. Aunque este tema es todavía novedoso, existen algunos estudios que demuestran su gran eficacia, puesto que permiten demostrar por qué las redes toman unas decisiones u otras en el diagnóstico de enfermedades. A su vez, todo ello contribuye a que los sistemas de DL, tales como las CNNs diseñadas para clasificar imágenes médicas, ganen la confianza por parte de los especialistas médicos. Por lo tanto, se espera que, gracias a XAI, estos modelos se aplicarán con mayor frecuencia en la práctica clínica (Deperlioglu et al., 2022; Samek et al., 2019).

En este capítulo final se presentarán las conclusiones y aportaciones más relevantes del TFM realizado. También, se expondrán algunas de las limitaciones que se han detectado y se propondrán varias líneas futuras con las que poder continuar el estudio.

6.2. Contribuciones originales

En este TFM se ha desarrollado un sistema de análisis automático de retinografías aplicado al diagnóstico de patologías oculares y se han aplicado técnicas XAI que permiten explicar los resultados. En los últimos años, el número de enfermedades que afectan al ojo humano han aumentado considerable y las predicciones de los

científicos determinan que van a seguir incrementándose en los próximos años. Entre las enfermedades que se han convertido en importantes causas de pérdida de visión a nivel mundial, se encuentran la RD, el glaucoma, las cataratas y la DMAE. Para su detección temprana, el análisis de retinografías es una de las técnicas más utilizadas por los oftalmólogos especialistas, puesto que se trata de un método no invasivo que permite una inspección clara de las principales estructuras oculares.

Para la implementación del método se ha empleado la arquitectura DenseNet-121, que permite obtener muy buenos resultados. También, se han aplicado 6 métodos de atribución con el fin de analizar las decisiones tomadas por la CNN. Entre las contribuciones más relevantes de este trabajo destacan las siguientes:

1. Adaptación del método propuesto en nuestro antiguo trabajo. Se ha hecho uso del mismo modelo de CNN como consecuencia del gran rendimiento que ofrece. No obstante, se han necesitado realizar pequeñas modificaciones sobre el mismo para poder posteriormente implementar los métodos XAI haciendo uso de las librerías disponibles en Python.

2. Utilización de una nueva base de datos. Con el objetivo de implementar un sistema automático que fuese capaz de generalizar un rango más amplio de patologías oculares, se ha hecho uso de una nueva BD pública (denominada RFMiD). Para ello, las imágenes correspondientes a su conjunto de entrenamiento y de validación, se han mezclado con las de la BD privada para llevar a cabo el entrenamiento del modelo.

3. Explicación e interpretación de los resultados mediante la aplicación de diferentes técnicas XAI de atribución. Concretamente, se han implementado seis tipos como son: *SHAP, Input x Gradient, Integrated Gradients, DeepTaylor, SmoothGrad y LRP−ε.* En la literatura actual, existen muy pocos estudios donde se apliquen estas técnicas para la tarea específica de cribado de patologías del ojo.

6.3. Conclusiones

La principal conclusión que se puede extraer de este trabajo es que los métodos de DL complementados con XAI, son muy útiles en la detección automática de la presencia de patología ocular, así como para entender las características de la imagen más relevantes para la ayuda al diagnóstico. La explicabilidad de los

resultados obtenidos en los algoritmos de DL está cobrando cada vez un mayor interés puesto que, permite reducir su carácter de "caja negra" y, por lo tanto, aumentar la confianza por parte de los médicos especialistas. En nuestro caso, se han utilizado métodos de atribución para entender las predicciones de un sistema automático de cribado de enfermedades oculares. A su vez, este sistema podría formar parte de una fase previa en el *screening* de la RD, que permitiría ahorrar tiempo y esfuerzos ya que evitaría el procesado posterior en aquellos casos en los que no se detectan signos patológicos. Por tanto, se cumple el objetivo de utilizar técnicas XAI en la detección de enfermedades tales como la RD, así como el resto de objetivos expuestos en el capítulo 1 de la memoria. A continuación, se detallan los objetivos y su grado de cumplimiento, así como las conclusiones más relevantes del trabajo.

1. Con el fin de entender todos los conceptos relacionados con el problema a abordar, se ha llevado a cabo una revisión bibliográfica exhaustiva sobre los diferentes métodos que existen para la clasificación automática de imágenes médicas, así como las técnicas XAI más utilizadas. En la actualidad, son las CNNs los modelos de DL más empleados para esta tarea específica, puesto que son las que mejores resultados obtienen. Respecto a XAI, los métodos de atribución son los más comunes debido a su fácil implementación.

2. Familiarización con las BBDD seleccionadas. La BD privada contiene 1000 retinografías pertenecientes a pacientes diabéticos que pueden presentar patologías oculares asociadas a su enfermedad, tales como la RD, el glaucoma o la DMAE. La BD pública (RFMiD) cuenta con 3200 imágenes de fondo de ojo pertenecientes a 46 enfermedades distintas, luego está enfocada a la creación de sistemas generalizables capaces de detectar tanto las patologías oculares más conocidas como las más raras, puesto que estas últimas también pueden afectar considerablemente a la vista. Para este propósito, se ha decidido utilizar ambas BBDD para crear una única con 4200 imágenes en total (2720 de entrenamiento, 740 de validación y 740 de test).

3. Implementación del método propuesto y de las diferentes técnicas XAI, para lo cual se ha utilizado el lenguaje Python junto con sus librerías TensorFlow, Keras, Innvestigate y SHAP. En los sistemas de DL resulta extremadamente difícil entender las conclusiones a las que llegan las redes neuronales. En este sentido, las técnicas XAI permiten comprender mejor las predicciones que realizan los modelos.

4. El funcionamiento del método se ha evaluado de manera exhaustiva utilizando

hasta cinco métricas diferentes. En la literatura actual, la mayoría de los métodos no incluyen todas estas métricas de evaluación por lo que, se puede decir que se ha comprobado con fiabilidad el desempeño del sistema desarrollado.

5. Se ha llevado a cabo la comparación entre las diferentes técnicas XAI implementadas para así poder comprobar cuál era la óptima.

Las conclusiones principales extraídas de este TFM son:

1. Las técnicas de preprocesado necesarias para las imágenes de entrada de un sistema de DL son muy básicas y sencillas puesto que estos sistemas ya funcionan muy bien de por sí, extrayendo las características directamente de las imágenes originales.

2. Aunque es cierto que el mundo del DL puede llegar a ser muy complejo, la implementación de un sistema automático y de los métodos XAI mediante Python es relativamente sencillo. Esto se debe a que dicho lenguaje cuenta con librerías que facilitan considerablemente el trabajo.

3. Los resultados obtenidos mediante la arquitectura DenseNet-121 muestran un elevado rendimiento. Además, la implementación de XAI mediante el uso de métodos de atribución en CNNs es relativamente sencilla y permite observar de manera clara por qué la red toma una decisión u otra. Entre todas las técnicas XAI que se han explorado, la mejor es IG puesto que es la que obtiene los mapas de atribución más fáciles de entender de cara al usuario.

4. Estos sistemas cuentan con la ventaja de la automatización, es decir una vez configurados con los parámetros óptimos, no necesitan la intervención humana para el análisis de imágenes. Por tanto, todo este sistema (CNN + XAI) se podría emplear en la práctica clínica, ya que mejoraría la atención primaria ahorrando tiempo, costes y carga de trabajo a los oftalmólogos especialistas, quienes gracias a XAI podrían entender mucho mejor los resultados obtenidos.

6.4. Limitaciones y líneas futuras

Aunque es cierto que los resultados que se han obtenido en este TFM son bastante

exitosos, el estudio también presenta ciertas limitaciones que cabe mencionar:

1. Elección de los parámetros de la red y coste computacional. El valor inicial para algunos hiperparámetros ha sido elegido de manera experimental, por lo que se podrían haber ejecutado más pruebas para perfeccionar estos valores. No obstante, el uso de CNNs conlleva un gran coste computacional ya que requieren la ejecución de un gran número de operaciones, así como el almacenamiento de una gran cantidad de datos.

2. Sobreentrenamiento del modelo. Al analizar el conjunto de entrenamiento y de validación, se ha podido comprobar que se obtienen mejores resultados en el primero respecto al segundo. Esto podría deberse a que el modelo se ha adaptado demasiado a los datos de entrenamiento y por eso, generaliza algo peor.

3. Tamaño de la BD empleada. Para poder detectar con mayor éxito múltiples tipos de patologías oculares, se necesitaría completar la BD con un mayor número de imágenes correspondientes a cada posible enfermedad. Además, también sería bueno incluir más imágenes reales obtenidas en la práctica clínica, es decir tomadas con diferentes retinógrafos, ángulos de visión y por distintos oftalmólogos. Todo ello contribuiría a que el método desarrollado pudiese ser más universal y, por lo tanto, que se pudiera utilizar con prácticamente todas las BBDD de retinografías accesibles públicamente. Al aumentar el número de imágenes, cabe esperar una mejora en los resultados ya que las CNNs requieren muchos datos para poder entrenarse y conseguir una generalización de calidad.

4. Desbalanceo en las clases de la BD. La BD pública empleada carece de clases balanceadas puesto que en todos los casos, la clase predominante es la que contiene las imágenes con patologías y la clase minoritaria, la correspondiente a retinografías de pacientes sanos. Aunque se ha tratado de minimizar este problema a la hora de entrenar la red, ésta tiende a favorecer las clases mayoritarias, lo que se traduce en una menor precisión en la clasificación de la clase menos predominante. Esta es la razón por la que se han obtenido peores resultados en la detección de imágenes de retina sin patologías.

A continuación, se presentan algunas líneas futuras de trabajo con las que se podría continuar, teniendo en cuenta las limitaciones anteriores.

1. Realizar diferentes pruebas variando el valor de los hiperparámetros con el

objetivo de optimizar más los resultados y de evitar el sobreentrenamiento del modelo. También, se podrían probar diferentes estrategias en la aplicación de las técnicas de optimización utilizadas (*transfer learning*, *fine tuning* y *dropout*).

2. Incrementar el número de imágenes de la BD y balancear sus clases, con el objetivo de poder utilizar el método desarrollado para detectar con calidad una gran variedad de patologías oculares, cada vez más presentes en la práctica clínica.

3. Realizar una comparación cuantitativa más exhaustiva entre las técnicas XAI implementadas. Se espera que un método de atribución produzca mapas de características similares cuando el entrenamiento del modelo se repite de forma idéntica. Por tanto, para poder calcular métricas de evaluación, se debería repetir el entrenamiento del modelo varias veces y, para cada iteración, analizar las diferencias entre los distintos mapas obtenidos con cada método XAI.

Referencias

Abeyagunasekera, S. H. P., Perera, Y., Chamara, K., Kaushalya, U., Sumathipala, P., & Senaweera, O. (2022). LISA : Enhance the explainability of medical images unifying current XAI techniques. *2022 IEEE 7th International Conference for Convergence in Technology, I2CT 2022*. https://doi.org/10.1109/I2CT54291.2022.9824840

Abramoff, M. D., Garvin, M. K., & Sonka, M. (2010). Retinal imaging and image analysis. *IEEE Reviews in Biomedical Engineering*, *3*, 169–208. https://doi.org/10.1109/RBME.2010.2084567

Acharya, R. U., Yu, W., Zhu, K., Nayak, J., Lim, T. C., & Chan, J. Y. (2009). Identification of Cataract and Post-cataract Surgery Optical Images Using Artificial Intelligence Techniques. *Journal of Medical Systems 2009 34:4*, *34*(4), 619–628. https://doi.org/10.1007/S10916-009-9275-8

Aggarwal, C. C. (2018). Neural Networks and Deep Learning. *Neural Networks and Deep Learning*. https://doi.org/10.1007/978-3-319-94463-0

Al-Jarrah, M. A., & Shatnawi, H. (2017). Non-proliferative diabetic retinopathy symptoms detection and classification using neural network. *Journal of Medical Engineering & Technology*, *41*(6), 498–505. https://doi.org/10.1080/03091902.2017.1358772

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montalvon, G., Samek, W., Müller, K. R., Dähne, S., & Kindermans, P.-J. (2019). iNNvestigate Neural Networks. *Journal of Machine Learning Research*, *20*, 1–8. https://doi.org/10.1371/JOURNAL.PONE.0130140

Alcalá, V., Maeda, V., Zanella, L. A., Valladares, A., Celaya, J. M., & Galván, C. E. (2020). Convolutional Neural Network for Classification of Diabetic Retinopathy Grade. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12468 LNAI*, 104–118. https://doi.org/10.1007/978-3-030-60884-2_8/COVER

Aliseda, D., & Berastegui, L. (2008). *Retinopatía diabética*. *3*, 23–34. https://scielo.isciii.es/scielo.php?script=sci_abstract&pid=S1137-66272008000600003

Allison, K., Patel, D., & Alabi, O. (2020). Epidemiology of Glaucoma: The Past, Present, and Predictions for the Future. *Cureus*, *12*(11). https://doi.org/10.7759/CUREUS.11686

Alyoubi, W. L., Shalash, W. M., & Abulkhair, M. F. (2020). Diabetic retinopathy detection through deep learning techniques: A review. *Informatics in Medicine Unlocked*, *20*, 100377. https://doi.org/10.1016/J.IMU.2020.100377

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures,

challenges, applications, future directions. *Journal of Big Data 2021 8:1*, *8*(1), 1–74. https://doi.org/10.1186/S40537-021-00444-8

American Academy of Ophthalmology. (2016, August). *Management of Traumatic Cataract*. https://www.aao.org/eyenet/article/management-of-traumatic-cataract

American Academy of Ophthalmology. (2022). *Retinal Artery Occlusion*. https://eyewiki.aao.org/Retinal_Artery_Occlusion

Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2017). Towards better understanding of gradient-based attribution methods for Deep Neural Networks. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. https://doi.org/10.48550/arxiv.1711.06104

Araújo, T., Aresta, G., Mendonça, L., Penas, S., Maia, C., Carneiro, Â., Mendonça, A. M., & Campilho, A. (2020). DR|GRADUATE: uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. *Medical Image Analysis*, *63*. https://doi.org/10.1016/j.media.2020.101715

Asbell, P. A., Dualan, I., Mindel, J., Brocks, D., Ahmad, M., & Epstein, S. (2005). Age-related cataract. *The Lancet*, *365*(9459), 599–609. https://doi.org/10.1016/S0140-6736(05)17911-2

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS ONE*, *10*(7). https://doi.org/10.1371/JOURNAL.PONE.0130140

Badar, M., Haris, M., & Fatima, A. (2020). Application of deep learning for retinal image analysis: A review. *Computer Science Review*, *35*. https://doi.org/10.1016/J.COSREV.2019.100203

Bae, J. P., Kim, K. G., Kang, H. C., Jeong, C. B., Park, K. H., & Hwang, J. M. (2011). A study on hemorrhage detection using hybrid method in fundus images. *Journal of Digital Imaging*, *24*(3), 394–404. https://doi.org/10.1007/S10278-010-9274-9

Baratloo, A., Hosseini, M., Negida, A., & Ashal, G. El. (2015). Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity. *Emergency*, *3*(2), 48. /pmc/articles/PMC4614595/

Baumal, C. R. (2018). Branch Retinal Artery Occlusion. *Atlas of Retinal OCT: Optical Coherence Tomography*, 96–97. https://doi.org/10.1016/B978-0-323-46121-4.00042-X

Beaser, R. S., Turell, W. A., & Howson, A. (2018). Strategies to Improve Prevention and Management in Diabetic Retinopathy: Qualitative Insights from a Mixed-Methods Study. *Diabetes Spectrum : A Publication of the American Diabetes Association*, *31*(1), 65. https://doi.org/10.2337/DS16-0043

Bell, S. J., Oluonye, N., Harding, P., & Moosajee, M. (2020). Congenital cataract: a guide to genetic and clinical management: *Https://Doi.Org/10.1177/2633004020938061*, *1*, 263300402093806. https://doi.org/10.1177/2633004020938061

Besenczi, R., Tóth, J., & Hajdu, A. (2016). A review on automatic analysis techniques for color fundus photographs. *Computational and Structural Biotechnology Journal*, *14*, 371–384. https://doi.org/10.1016/J.CSBJ.2016.10.001

Bianco, S., Cadene, R., Celona, L., & Napoletano, P. (2018). Benchmark analysis of representative deep neural network architectures. *IEEE Access*, *6*, 64270–64277. https://doi.org/10.1109/ACCESS.2018.2877890

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. https://dl.acm.org/doi/book/10.5555/525960

Bixler, J. E. (2019). Cataracts and Their Treatment in People with Diabetes. *Prevention and Management of Diabetes-Related Eye Disease*. https://doi.org/10.2337/DB20191-6

Blair, K., & Czyz, C. N. (2021). *Central Retinal Vein Occlusion*. StatPearls. https://www.ncbi.nlm.nih.gov/books/NBK525985/

Blair K, & Czyz C. N. (2020). *Retinal Detachment*. StatPearls. https://www.ncbi.nlm.nih.gov/books/NBK551502/

Bourdon, P., Ahmed, O. Ben, Urruty, T., Djemal, K., & Fernandez-Maloigne, C. (2021). Explainable AI for Medical Imaging: Knowledge Matters. *Multi-Faceted Deep Learning*, 267–292. https://doi.org/10.1007/978-3-030-74478-6_11

Brownlee, J. (2018). *What is the Difference Between a Batch and an Epoch in a Neural Network?* https://machinelearningmastery.com/difference-between-a-batch-and-an-epoch/

Cen, L. P., Ji, J., Lin, J. W., Ju, S. T., Lin, H. J., Li, T. P., Wang, Y., Yang, J. F., Liu, Y. F., Tan, S., Tan, L., Li, D., Wang, Y., Zheng, D., Xiong, Y., Wu, H., Jiang, J., Wu, Z., Huang, D., … Zhang, M. (2021). Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nature Communications 2021 12:1*, *12*(1), 1–13. https://doi.org/10.1038/s41467-021-25138-w

Chodick, G., Bekiroglu, N., Hauptmann, M., Alexander, B. H., Freedman, D. M., Doody, M. M., Cheung, L. C., Simon, S. L., Weinstock, R. M., Bouville, A., & Sigurdson, A. J. (2008). Risk of Cataract after Exposure to Low Doses of Ionizing Radiation: A 20-Year Prospective Cohort Study among US Radiologic Technologists. *American Journal of Epidemiology*, *168*(6), 620. https://doi.org/10.1093/AJE/KWN171

Choi, J. Y., Yoo, T. K., Seo, J. G., Kwak, J., Um, T. T., & Rim, T. H. (2017). Multi-categorical deep learning neural network to classify retinal images: A pilot study employing small database. *PloS One*, *12*(11). https://doi.org/10.1371/JOURNAL.PONE.0187336

Cilimkovic, M. (2015). *Neural Networks and Back Propagation Algorithm*.

Cochran, M. L., Mahabadi, N., & Czyz, C. N. (2018). *Branch Retinal Vein Occlusion*. StatPearls. https://www.ncbi.nlm.nih.gov/books/NBK535370/

Cole, E. D., Novais, E. A., Louzada, R. N., & Waheed, N. K. (2016). Contemporary retinal imaging techniques in diabetic retinopathy: a review. *Clinical & Experimental Ophthalmology*, *44*(4), 289–299. https://doi.org/10.1111/CEO.12711

Coney, J. M. (2019). Addressing unmet needs in diabetic retinopathy. *The American Journal of Managed Care* , *25*.

https://www.researchgate.net/publication/337424873_Addressing_unmet_needs_in_diabetic_retinopathy

Das, D., Biswas, S. K., & Bandyopadhyay, S. (2022). A critical review on diagnosis of diabetic retinopathy using machine learning and deep learning. *Multimedia Tools and Applications 2022*, 1–43. https://doi.org/10.1007/S11042-022-12642-4

Das, S., & Saha, S. K. (2022). Diabetic retinopathy detection and classification using CNN tuned by genetic algorithm. *Multimedia Tools and Applications 2022 81:6*, *81*(6), 8007–8020. https://doi.org/10.1007/S11042-021-11824-W

Dashtbozorg, B., Zhang, J., Huang, F., & Romeny, B. M. T. H. (2018). Retinal Microaneurysms Detection Using Local Convergence Index Features. *IEEE Transactions on Image Processing*, *27*(7), 3300–3315. https://doi.org/10.1109/TIP.2018.2815345

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2010). *ImageNet: A large-scale hierarchical image database*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Deperlioglu, O., Kose, U., Gupta, D., Khanna, A., Giampaolo, F., & Fortino, G. (2022). Explainable framework for Glaucoma diagnosis by image processing and convolutional neural network synergy: Analysis with doctor evaluation. *Future Generation Computer Systems*, *129*, 152–169. https://doi.org/10.1016/J.FUTURE.2021.11.018

Diaz, A., Morales, S., Naranjo, V., Köhler, T., Mossi, J. M., & Navea, A. (2019). CNNs for automatic glaucoma assessment using fundus images: An extensive validation. *BioMedical Engineering Online*, *18*(1). https://doi.org/10.1186/S12938-019-0649-Y

Doshi, D., Shenoy, A., Sidhpura, D., & Gharpure, P. (2017). Diabetic retinopathy detection using deep convolutional neural networks. *International Conference on Computing, Analytics and Security Trends, CAST 2016*, 261–266. https://doi.org/10.1109/CAST.2016.7914977

Early Treatment Diabetic Retinopathy Study Research Group. (1991). Grading Diabetic Retinopathy from Stereoscopic Color Fundus Photographs — An Extension of the Modified Airlie House Classification. ETDRS Report Number 10. *Ophthalmology*, *98*(5), 786–806. https://doi.org/10.1016/j.ophtha.2020.01.030

Emmert-Streib, F., Yang, Z., Feng, H., Tripathi, S., & Dehmer, M. (2020). An Introductory Review of Deep Learning for Prediction Models With Big Data. *Frontiers in Artificial Intelligence*, *3*. https://doi.org/10.3389/FRAI.2020.00004

Farris, W., & Waymack, J. R. (2021). *Central Retinal Artery Occlusion* . https://pubmed.ncbi.nlm.nih.gov/29262124/

Fernández Revuelta, A. (2012). *Técnica de exploración del fondo del ojo*. 5. https://medicinainternaaldia.files.wordpress.com/2014/12/semiologc3ada-fondo-del-ojo.pdf

Ferris, F. L., Podgor, M. J., & Davis, M. D. (1987). Macular Edema in Diabetic Retinopathy Study Patients: Diabetic Retinopathy Study Report Number 12. *Ophthalmology*, *94*(7), 754–760. https://doi.org/10.1016/S0161-6420(87)33526-2

Ganguly, D., Chakraborty, S., Balitanas, M., & Kim, T. H. (2010). Medical imaging: A review. *Communications in Computer and Information Science*, *78 CCIS*, 504–516. https://doi.org/10.1007/978-3-642-16444-6_63

García Gadañón, M. (2008). *Procesado de retinografías basado en redes neuronales para la detección automática de lesiones asociadas a la retinopatía diabética* [Universidad de Valladolid]. https://www.worldcat.org/title/procesado-de-retinografias-basado-en-redes-neuronales-para-la-deteccion-automatica-de-lesiones-asociadas-a-la-retinopatia-diabetica/oclc/630813211

Gayathri, S., Gopi, V. P., & Palanisamy, P. (2020). A lightweight CNN for Diabetic Retinopathy classification from fundus images. *Biomedical Signal Processing and Control*, *62*, 102115. https://doi.org/10.1016/J.BSPC.2020.102115

Ghazi, N. G., & Green, W. R. (2002). Pathology and pathogenesis of retinal detachment. *Eye 2002 16:4*, *16*(4), 411–421. https://doi.org/10.1038/sj.eye.6700197

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Chapter 9. Deep Learning*. MIT Press. https://mitpress.mit.edu/9780262035613/

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., & Chen, T. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, *77*, 354–377. https://doi.org/10.1016/J.PATCOG.2017.10.013

Haglin, J. M., Jimenez, G., & Eltorai, A. E. M. (2019). Artificial neural networks in medicine. *Health and Technology*, *9*(1), 1–6. https://doi.org/10.1007/S12553-018-0244-4/FIGURES/3

Hanúsková, V., Pavlovićová, J., Oravec, M., & Blaško, R. (2013). Diabetic rethinopathy screening by bright lesions extraction from fundus images. *Journal of Electrical Engineering*, *64*(5), 311–316. https://doi.org/10.2478/JEE-2013-0045

Harasymowycz, P., Birt, C., Gooi, P., Heckler, L., Hutnik, C., Jinapriya, D., Shuba, L., Yan, D., & Day, R. (2016). Medical Management of Glaucoma in the 21st Century from a Canadian Perspective. *Journal of Ophthalmology*, *2016*. https://doi.org/10.1155/2016/6509809

Hossain, M. R., Afroze, S., Siddique, N., & Hoque, M. M. (2020). Automatic Detection of Eye Cataract using Deep Convolution Neural Networks (DCNNs). *2020 IEEE Region 10 Symposium, TENSYMP 2020*, 1333–1338. https://doi.org/10.1109/TENSYMP50017.2020.9231045

Hua, C. H., Huynh-The, T., & Lee, S. (2020). DRAN: Densely Reversed Attention based Convolutional Network for Diabetic Retinopathy Detection. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, *2020-July*, 1992–1995. https://doi.org/10.1109/EMBC44109.2020.9175355

Hussain, F., Hussain, R., & Hossain, E. (2021). *Explainable Artificial Intelligence (XAI): An Engineering Perspective*. https://cutt.ly/nh8rMyj

International Diabetes Federation. (2020, May). *Diabetes Complications*. https://www.idf.org/aboutdiabetes/complications.html

International Diabetes Federation. (2021). *IDF Diabetes Atlas 2021* . https://diabetesatlas.org/

Ishtiaq, U., Abdul Kareem, S., Abdullah, E. R. M. F., Mujtaba, G., Jahangir, R., & Ghafoor, H. Y. (2019). Diabetic retinopathy detection through artificial intelligent techniques: a review and open issues. *Multimedia Tools and Applications 2019 79:21*, *79*(21), 15209–15252. https://doi.org/10.1007/S11042-018-7044-8

Islam, M. M., Yang, H. C., Poly, T. N., Jian, W. S., & (Jack) Li, Y. C. (2020). Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: A systematic review and meta-analysis. *Computer Methods and Programs in Biomedicine*, *191*, 105320. https://doi.org/10.1016/J.CMPB.2020.105320

Issac, A., Partha Sarathi, M., & Dutta, M. K. (2015). An adaptive threshold based image processing technique for improved glaucoma detection and classification. *Computer Methods and Programs in Biomedicine*, *122*(2), 229–244. https://doi.org/10.1016/J.CMPB.2015.08.002

Ivanovs, M., Kadikis, R., & Ozols, K. (2021). Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, *150*, 228–234. https://doi.org/10.1016/J.PATREC.2021.06.030

Jaffe, G. J., Ying, G. S., Toth, C. A., Daniel, E., Grunwald, J. E., Martin, D. F., & Maguire, M. G. (2019). Macular Morphology and Visual Acuity in Year Five of the Comparison of Age-related Macular Degeneration Treatments Trials. *Ophthalmology*, *126*(2), 252–260. https://doi.org/10.1016/J.OPHTHA.2018.08.035

Javadi, M. A., & Zarei-Ghanavati, S. (2008). Cataracts in Diabetic Patients: A Review Article. *Journal of Ophthalmic & Vision Research*, *3*(1), 52. /pmc/articles/PMC3589218/

Jiang, H., Yang, K., Gao, M., Zhang, D., Ma, H., & Qian, W. (2019). An Interpretable Ensemble Deep Learning Model for Diabetic Retinopathy Disease Classification. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, *2019*, 2045–2048. https://doi.org/10.1109/EMBC.2019.8857160

Johnson, N. S., Vulimiri, P. S., To, A. C., Zhang, X., Brice, C. A., Kappes, B. B., & Stebner, A. P. (2020). Machine learning for materials developments in metals additive manufacturing. *Additive Manufacturing*, *36*, 101641. https://doi.org/10.1016/J.ADDMA.2020.101641

Joshi, S., & Karule, P. T. (2018). A review on exudates detection methods for diabetic retinopathy. *Biomedicine & Pharmacotherapy*, *97*, 1454–1460. https://doi.org/10.1016/J.BIOPHA.2017.11.009

Karia, N. (2010). Retinal vein occlusion: pathophysiology and treatment options. *Clinical Ophthalmology (Auckland, N.Z.)*, *4*(1), 809. https://doi.org/10.2147/OPTH.S7631

Keane, P. A., & Sadda, S. R. (2014). Retinal Imaging in the Twenty-First Century: State of the Art and Future Directions. *Ophthalmology*, *121*(12), 2489–2500. https://doi.org/10.1016/J.OPHTHA.2014.07.054

Keskar, N. S., & Socher, R. (2017). *Improving Generalization Performance by Switching from Adam to*

SGD. https://doi.org/10.48550/arxiv.1712.07628

Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review 2020 53:8*, *53*(8), 5455–5516. https://doi.org/10.1007/S10462-020-09825-6

Kharroubi, A. T., & Darwish, H. M. (2015). Diabetes mellitus: The epidemic of the century. *World Journal of Diabetes*, *6*(6), 850. https://doi.org/10.4239/WJD.V6.I6.850

Kim, M., Yun, J., Cho, Y., Shin, K., Jang, R., Bae, H. J., & Kim, N. (2019). Deep Learning in Medical Imaging. *Neurospine*, *16*(4), 657. https://doi.org/10.14245/NS.1938396.198

Kolhe, S., Guru, S. K., & Student, P. G. (2007). Remote Automated Cataract Detection System Based on Fundus Images. *International Journal of Innovative Research in Science, Engineering and Technology (An ISO*, *3297*. https://doi.org/10.15680/IJIRSET.2015.0506152

Krause, J., Gulshan, V., Rahimy, E., Karth, P., Widner, K., Corrado, G. S., Peng, L., & Webster, D. R. (2017). Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*, *125*(8), 1264–1272. https://doi.org/10.1016/j.ophtha.2018.01.034

Kumar, S., Pathak, S., & Kumar, B. (2019). Automated detection of eye related diseases using digital image processing. *Handbook of Multimedia Information Security: Techniques and Applications*, 513–544. https://doi.org/10.1007/978-3-030-15887-3_25/COVER

Kwasigroch, A., Jarzembinski, B., & Grochowski, M. (2018). Deep CNN based decision support system for detection and assessing the stage of diabetic retinopathy. *2018 International Interdisciplinary PhD Workshop, IIPhDW 2018*, 111–116. https://doi.org/10.1109/IIPHDW.2018.8388337

Lambert-Cheatham, N., Jusufbegovic, D., & Corson, T. W. (2022). Intraocular and Orbital Cancers. *Comprehensive Pharmacology*, 146–193. https://doi.org/10.1016/B978-0-12-820472-6.00024-4

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2323. https://doi.org/10.1109/5.726791

Li, J. Q., Welchowski, T., Schmid, M., Letow, J., Wolpers, A. C., Holz, F. G., & Finger, R. P. (2018). Retinal Diseases in Europe. *Prevalence, Incidence and Healthcare Needs*. https://miloftalmica.it/wp-content/uploads/2021/07/Euretina-Retinal-Diseases.pdf

Lim, L. S., Mitchell, P., Seddon, J. M., Holz, F. G., & Wong, T. Y. (2012). Age-related macular degeneration. *The Lancet*, *379*(9827), 1728–1738. https://doi.org/10.1016/S0140-6736(12)60282-7

Liu, W., Chen, L., & Chen, Y. (2018). Age Classification Using Convolutional Neural Networks with the Multi-class Focal Loss. *IOP Conference Series: Materials Science and Engineering*, *428*(1), 012043. https://doi.org/10.1088/1757-899X/428/1/012043

Lopez Pinaya, W. H., Vieira, S., Garcia-Dias, R., & Mechelli, A. (2019). Convolutional neural networks. *Machine Learning: Methods and Applications to Brain Disorders*, 173–191. https://doi.org/10.1016/B978-0-12-815739-8.00010-9

Lu, C.-K., Tang, T. B., Laude, A., Dhillon, B., & Murray, A. F. (2012). Parapapillary atrophy and optic disc region assessment (PANDORA): retinal imaging tool for assessment of the optic disc and parapapillary atrophy. *Https://Doi.Org/10.1117/1.JBO.17.10.106010*, *17*(10), 106010. https://doi.org/10.1117/1.JBO.17.10.106010

Lu, L., Ren, P., Lu, Q., Zhou, E., Yu, W., Huang, J., He, X., & Han, W. (2021). Analyzing fundus images to detect diabetic retinopathy (DR) using deep learning system in the Yangtze River delta region of China. *Annals of Translational Medicine*, *9*(3), 226–226. https://doi.org/10.21037/ATM-20-3275

Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *2017-December*, 4766–4775. https://doi.org/10.48550/arxiv.1705.07874

Luo, X., Li, J., Chen, M., Yang, X., & Li, X. (2021). Ophthalmic Disease Detection via Deep Learning with a Novel Mixture Loss Function. *IEEE Journal of Biomedical and Health Informatics*, *25*(9), 3332–3339. https://doi.org/10.1109/JBHI.2021.3083605

Mac Grory, B., Schrag, M., Biousse, V., Furie, K. L., Gerhard-Herman, M., Lavin, P. J., Sobrin, L., Tjoumakaris, S. I., Weyand, C. M., & Yaghi, S. (2021). Management of central retinal artery occlusion: A scientific statement from the American heart association. *Stroke*, *52*, E282–E294. https://doi.org/10.1161/STR.0000000000000366

Mahabadi Navid, & Al Khalili, Y. (2021). *Neuroanatomy, Retina* . StatPearls. https://pubmed.ncbi.nlm.nih.gov/31424894/

Matsushima, H., Iwamoto, H., Mukai, K., Katsuki, Y., Nagata, M., & Senoo, T. (2008). Preventing secondary cataract and anterior capsule contraction by modification of intraocular lenses. *Expert Rev. Med. Devices*, *5*(2), 197–207. https://doi.org/10.1586/17434440.5.2.197

McMonnies, C. W. (2017). Glaucoma history and risk factors. *Journal of Optometry*, *10*(2), 71–78. https://doi.org/10.1016/J.OPTOM.2016.02.003

Minarno, A. E., Mandiri, M. H. C., Azhar, Y., Bimantoro, F., Nugroho, H. A., & Ibrahim, Z. (2022). Classification of Diabetic Retinopathy Disease Using Convolutional Neural Network. *JOIV : International Journal on Informatics Visualization*, *6*(1), 12–18. https://doi.org/10.30630/JOIV.6.1.857

Mishra, S., Hanchate, S., & Saquib, Z. (2020). Diabetic retinopathy detection using deep learning. *Proceedings of the International Conference on Smart Technologies in Computing, Electrical and Electronics, ICSTCEE 2020*, 515–520. https://doi.org/10.1109/ICSTCEE49637.2020.9277506

Mittal, P., & Bhatnagar, C. (2021). Retinal Disease Classification Using Convolutional Neural Networks Algorithm. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(11), 5681–5689. https://www.turcomat.org/index.php/turkbilmat/article/view/6822

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K. R. (2017). Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, *65*, 211–222. https://doi.org/10.1016/J.PATCOG.2016.11.008

Muddamsetty, S. M., Jahromi, M. N. S., & Moeslund, T. B. (2021). Expert Level Evaluations for Explainable AI (XAI) Methods in the Medical Domain. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12663 LNCS*, 35–46. https://doi.org/10.1007/978-3-030-68796-0_3/COVER

Muñoz Zamarro, C. P. (2020). *Aplicación de técnicas de procesado de imagen y Deep Learning para la clasificación de imágenes de fondo de ojo en retinopatía diabética*.

Musat, O., Cernat, C., Labib, M., Gheorghe, A., Toma, O., Zamfir, M., & Boureanu, A. M. ari. (2015). DIABETIC MACULAR EDEMA. *Romanian Journal of Ophthalmology*, *59*(3), 133. https://doi.org/10.3126/nepjoph.v7i2.14956

Nagpal, D., Panda, S. N., Malarvel, M., Pattanaik, P. A., & Zubair Khan, M. (2021). A review of diabetic retinopathy: Datasets, approaches, evaluation metrics and future trends. *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1016/J.JKSUCI.2021.06.006

National Eye Institute. (2017, May). *Diabetic Eye Disease*. https://www.niddk.nih.gov/health-information/diabetes/overview/preventing-problems/diabetic-eye-disease

National Eye Institute. (2021). *Age-Related Macular Degeneration (AMD)* . 06. https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/age-related-macular-degeneration

National Eye Institute. (2022a, April). *Glaucoma* . https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/glaucoma

National Eye Institute. (2022b, April). *Retinal Detachment* . https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/retinal-detachment

National Eye Institute. (2022c, April 21). *Cataracts*. https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/cataracts

National Institutes of Health. (2016). *Symptoms & Causes of Diabetes*. https://www.niddk.nih.gov/health-information/diabetes/overview/symptoms-causes#symptoms

Nayak, J. (2013). Automated Classification of Normal, Cataract and Post Cataract Optical Eye Images using SVM Classifier. *Proceedings of the World Congress on Engineering and Computer Science 2013*, 23–25. http://www.iaeng.org/publication/WCECS2013/WCECS2013_pp542-545.pdf

Nayak, J., Acharya U., R., Bhat, P. S., Shetty, N., & Lim, T. C. (2009). Automated diagnosis of glaucoma using digital fundus images. *Journal of Medical Systems*, *33*(5), 337–346. https://doi.org/10.1007/S10916-008-9195-Z

Nayak, J., Bhat, P. S., Acharya U, R., Lim, C. M., & Kagathi, M. (2008). Automated identification of diabetic retinopathy stages using digital fundus images. *Journal of Medical Systems*, *32*(2), 107–115. https://doi.org/10.1007/S10916-007-9113-9

Nedjar, I., Brahimi, M., Mahmoudi, S., Abi-Ayad, K., & Chikh, M. A. (2022). *Interpretation of breast tumors classification using convolutional neural network visualization*. http://dspace.univ-

eloued.dz:80/xmlui/handle/123456789/10811

Nentwich, M. M., & Ulbig, M. W. (2015). Diabetic retinopathy - ocular complications of diabetes mellitus. *World Journal of Diabetes*, *6*(3), 489. https://doi.org/10.4239/WJD.V6.I3.489

Nguyen, Q. H., Muthuraman, R., Singh, L., Sen, G., Tran, A. C., Nguyen, B. P., & Chua, M. (2020). Diabetic retinopathy detection using deep learning. *ACM International Conference Proceeding Series*, 103–107. https://doi.org/10.1145/3380688.3380709

Nizami, A. A., & Gulani, A. C. (2019). Cataract. *StatPearls*. http://europepmc.org/books/NBK539699

O'Shea, K., & Nash, R. (2015). *An Introduction to Convolutional Neural Networks*. https://doi.org/10.48550/arxiv.1511.08458

Oh, E., Yoo, T. K., & Hong, S. (2015). Artificial Neural Network Approach for Differentiating Open-Angle Glaucoma From Glaucoma Suspect Without a Visual Field Test. *Investigative Ophthalmology & Visual Science*, *56*(6), 3957–3966. https://doi.org/10.1167/IOVS.15-16805

Oh, K., Kang, H. M., Leem, D., Lee, H., Seo, K. Y., & Yoon, S. (2021). Early detection of diabetic retinopathy based on deep learning and ultra-wide-field fundus images. *Scientific Reports 2021 11:1*, *11*(1), 1–9. https://doi.org/10.1038/s41598-021-81539-3

Pachade, S., Porwal, P., Thulkar, D., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., Giancardo, L., Quellec, G., & Mériaudeau, F. (2021). Retinal fundus multi-disease image dataset (Rfmid): A dataset for multi-disease detection research. *Data*, *6*(2), 1–14. https://doi.org/10.3390/DATA6020014

Paing, M. P., Choomchuay, S., & Rapeeporn Yodprom, M. D. (2017). Detection of lesions and classification of diabetic retinopathy using fundus images. *BMEiCON 2016 - 9th Biomedical Engineering International Conference*. https://doi.org/10.1109/BMEICON.2016.7859642

Pashaei, E., & Pashaei, E. (2021). Training Feedforward Neural Network Using Enhanced Black Hole Algorithm: A Case Study on COVID-19 Related ACE2 Gene Expression Classification. *Arabian Journal for Science and Engineering*, *46*(4), 3807–3828. https://doi.org/10.1007/S13369-020-05217-8/FIGURES/13

Phasuk, S., Poopresert, P., Yaemsuk, A., Suvannachart, P., Itthipanichpong, R., Chansangpetch, S., Manassakorn, A., Tantisevi, V., Rojanapongpun, P., & Tantibundhit, C. (2019). Automated Glaucoma Screening from Retinal Fundus Image Using Deep Learning. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 904–907. https://doi.org/10.1109/EMBC.2019.8857136

Pratap, T., & Kokil, P. (2019). Computer-aided diagnosis of cataract using deep transfer learning. *Biomedical Signal Processing and Control*, *53*. https://doi.org/10.1016/J.BSPC.2019.04.010

Priyanka Agnihotri, Smriti Gupta, S. L. C. S. K. S. (2018). Retinal Vascular Occlusive Disorders - Risk Factors. *International Journal of Innovative Research in Medical Science*, *3*(02). https://doi.org/10.23958/IJIRMS/VOL03-I02/07

Purola, P. K. M., Nättinen, J. E., Ojamo, M. U. I., Koskinen, S. V. P., Rissanen, H. A., Sainio, P. R. J., &

Uusitalo, H. M. T. (2021). Prevalence and 11-year incidence of common eye diseases and their relation to health-related quality of life, mental health, and visual impairment. *Quality of Life Research*, *30*(8), 2311–2327. https://doi.org/10.1007/S11136-021-02817-1/FIGURES/7

Quellec, G., Lamard, M., Conze, P. H., Massin, P., & Cochener, B. (2020). Automatic detection of rare pathologies in fundus photographs using few-shot learning. *Medical Image Analysis*, *61*, 101660. https://doi.org/10.1016/J.MEDIA.2020.101660

Ravudu, M., Jain, V., & Kunda, M. M. R. (2012). Review of image processing techniques for automatic detection of eye diseases. *Proceedings of the International Conference on Sensing Technology, ICST*, 320–325. https://doi.org/10.1109/ICSENST.2012.6461695

Rawat, W., & Wang, Z. (2017). Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Computation*, *29*(9), 2352–2449. https://doi.org/10.1162/NECO_A_00990

Retina - Anatomía del ojo - Visión. Definiciones y conceptos. (2018). https://definicionesyconceptos.com/retina-anatomia-del-ojo-vision/

Riordan-Eva, P. (2012). *Vaughan y Asbury. Oftalmología general* (18A ed.). https://books.google.com/books/about/Vaughan_y_Asbury_oftalmología_general_1.html?hl=es&id=_kiNCgAAQBAJ

Ritter, F., Boskamp, T., Homeyer, A., Laue, H., Schwier, M., Link, F., & Peitgen, H. O. (2011). Medical image analysis. *IEEE Pulse*, *2*(6), 60–70. https://doi.org/10.1109/MPUL.2011.942929

Romero-Oraá, R., García, M., Oraá-Pérez, J., López, M. I., & Hornero, R. (2019). Transfer learning para evaluar de forma automática la calidad en imágenes de fondo de ojo. *XXXVII Congreso Anual de La Sociedad Española de Ingeniería Biomédica (Caseib 2019)*. http://www.gib.tel.uva.es/simplicity/docs/5de7acb208907_caseib_2019_RRO_r.pdf

Saeed, F., Hussain, M., & Aboalsamh, H. A. (2021). Automatic Diabetic Retinopathy Diagnosis Using Adaptive Fine-Tuned Convolutional Neural Network. *IEEE Access*, *9*, 41344–41359. https://doi.org/10.1109/ACCESS.2021.3065273

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (2019). *Explainable AI : interpreting, explaining and visualizing deep learning* (1st ed.). Springer Cham.

Sarki, R., Ahmed, K., Wang, H., & Zhang, Y. (2020). Automated detection of mild and multi-class diabetic eye diseases using deep learning. *Health Information Science and Systems 2020 8:1*, *8*(1), 1–9. https://doi.org/10.1007/S13755-020-00125-5

Sarki, R., Ahmed, K., Wang, H., Zhang, Y., & Wang, K. (2022). Convolutional Neural Network for Multi-class Classification of Diabetic Eye Disease. *EAI Endorsed Transactions on Scalable Information Systems*, *9*(4), e5–e5. https://doi.org/10.4108/EAI.16-12-2021.172436

Sayres, R., Taly, A., Rahimy, E., Blumer, K., Coz, D., Hammel, N., Krause, J., Narayanaswamy, A., Rastegar, Z., Wu, D., Xu, S., Barb, S., Joseph, A., Shumski, M., Smith, J., Sood, A. B., Corrado, G. S., Peng, L., & Webster, D. R. (2019). Using a Deep Learning Algorithm and Integrated Gradients

Explanation to Assist Grading for Diabetic Retinopathy. *Ophthalmology*, *126*(4), 552–564. https://doi.org/10.1016/J.OPHTHA.2018.11.016

Schmidt-Erfurth, U., Garcia-Arumi, J., Bandello, F., Berg, K., Chakravarthy, U., Gerendas, B. S., Jonas, J., Larsen, M., Tadayoni, R., & Loewenstein, A. (2017). Guidelines for the Management of Diabetic Macular Edema by the European Society of Retina Specialists (EURETINA). *Ophthalmologica*, *237*(4), 185–222. https://doi.org/10.1159/000458539

Seçkin Ayhan, M., Benedikt Kümmerle, L., Kühlewein, L., Inhoffen, W., Aliyeva, G., Ziemssen, F., & Berens, P. (2021). *Clinical Validation of Saliency Maps for Understanding Deep Neural Networks in Ophthalmology*. https://doi.org/10.1101/2021.05.05.21256683

Seoud, L., Hurtut, T., Chelbi, J., Cheriet, F., & Langlois, J. M. P. (2016). Red Lesion Detection Using Dynamic Shape Features for Diabetic Retinopathy Screening. *IEEE Transactions on Medical Imaging*, *35*(4), 1116–1126. https://doi.org/10.1109/TMI.2015.2509785

Serener, A., & Serte, S. (2019). Transfer learning for early and advanced glaucoma detection with convolutional neural networks. *TIPTEKNO 2019 - Tip Teknolojileri Kongresi*. https://doi.org/10.1109/TIPTEKNO.2019.8894965

Setiawan, W., & Damayanti, F. (2020). Layers Modification of Convolutional Neural Network for Pneumonia Detection. *Journal of Physics: Conference Series*, *1477*(5). https://doi.org/10.1088/1742-6596/1477/5/052055

SHAP documentation. (2018). *shap.DeepExplainer* . https://shap-lrjball.readthedocs.io/en/latest/generated/shap.DeepExplainer.html

Shehzad, M., Qadri, S., Aslam, T., Furqan Qadri, S., Razzaq, A., Shah Muhammad, S., Ali Nawaz, S., & Ahmad, N. (2020). Machine Vision Based Identification of Eye Cataract Stages Using Texture Features. *Life Sci J*, *17*(8), 44–50. https://doi.org/10.7537/marslsj170820.07

Shrikumar, A., Greenside, P., Shcherbina, A. Y., & Kundaje, A. (2016). Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. *ArXiv*, *1*, 0–5. https://doi.org/10.48550/arxiv.1605.01713

Sikder, N., Masud, M., Bairagi, A. K., Arif, A. S. M., Nahid, A. Al, & Alhumyani, H. A. (2021). Severity Classification of Diabetic Retinopathy Using an Ensemble Learning Algorithm through Analyzing Retinal Images. *Symmetry 2021, Vol. 13, Page 670*, *13*(4), 670. https://doi.org/10.3390/SYM13040670

Simanjuntak, R. B. J., Fu'Adah, Y., Magdalena, R., Saidah, S., Wiratama, A. B., & Da'Wan Salim Ubaidah, I. (2022). Cataract Classification Based on Fundus Images Using Convolutional Neural Network. *JOIV : International Journal on Informatics Visualization*, *6*(1), 33–38. https://doi.org/10.30630/JOIV.6.1.856

Singh, A., Mohammed, A. R., Zelek, J., & Lakshminarayanan, V. (2020). Interpretation of deep learning using attributions: application to ophthalmic diagnosis. *Https://Doi.Org/10.1117/12.2568631*, *11511*, 39–49. https://doi.org/10.1117/12.2568631

Singh, A., Sengupta, S., J, J. B., Mohammed, A. R., Faruq, I., Jayakumar, V., Zelek, J., & Lakshminarayanan, V. (2020). What is the Optimal Attribution Method for Explainable Ophthalmic Disease Classification? *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12069 LNCS*, 21–31. https://doi.org/10.1007/978-3-030-63419-3_3/COVER

Singh, A., Sengupta, S., & Lakshminarayanan, V. (2020). *Explainable deep learning models in medical image analysis*.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). *SmoothGrad: removing noise by adding noise*. https://doi.org/10.48550/arxiv.1706.03825

Smith, L. N. (2018). *A disciplined approach to neural network hyper-parameters: Part 1 -- learning rate, batch size, momentum, and weight decay*. https://doi.org/10.48550/arxiv.1803.09820

Snell, R. S. (2007). *Neuroanatomía clínica* (6a ed.). Panamericana. https://books.google.es/books?hl=es&lr=&id=9AjM5_4tmMkC&oi=fnd&pg=PR7&dq=Neuroanatomía+Clinica+retina&ots=XqASAduTI1&sig=F9B8OjrOflCCyWahZYHauvQ3cyc#v=onepage&q=Neuroanatomía Clinica retina&f=false

Soltani, A., Badaoui, A., Battikh, T., & Jabri, I. (2018). A Novel System for Glaucoma Diagnosis Using Artificial Neural Network Classification. *2018 5th International Conference on Control, Decision and Information Technologies, CoDIT 2018*, 1128–1133. https://doi.org/10.1109/CODIT.2018.8394940

Spina, C. La, Benedetto, U. De, Parodi, M. B., Coscas, G., & Bandello, F. (2012). Practical Management of Retinal Vein Occlusions. *Ophthalmology and Therapy*, *1*(1). https://doi.org/10.1007/S40123-012-0003-Y

Srivastava, S. K., & Fekrat, S. (2006). Venous Obstructive Disease. *Retinal Imaging*, 248–260. https://doi.org/10.1016/B978-0-323-02346-7.50028-4

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic Attribution for Deep Networks. *34th International Conference on Machine Learning, ICML 2017*, *7*, 5109–5118. https://doi.org/10.48550/arxiv.1703.01365

Tawfik, H. R. M., Birry, R. A. K., & Saad, A. A. (2018). Early Recognition and Grading of Cataract Using a Combined Log Gabor/Discrete Wavelet Transform with ANN and SVM. *Engineering and Technology International Journal of Computer and Information Engineering*, *12*(12). https://publications.waset.org/10009852/early-recognition-and-grading-of-cataract-using-a-combined-log-gabordiscrete-wavelet-transform-with-ann-and-svm

Teo, Z. L., Tham, Y. C., Yu, M., Chee, M. L., Rim, T. H., Cheung, N., Bikbov, M. M., Wang, Y. X., Tang, Y., Lu, Y., Wong, I. Y., Ting, D. S. W., Tan, G. S. W., Jonas, J. B., Sabanayagam, C., Wong, T. Y., & Cheng, C. Y. (2021). Global Prevalence of Diabetic Retinopathy and Projection of Burden through 2045: Systematic Review and Meta-analysis. *Ophthalmology*, *128*(11), 1580–1591. https://doi.org/10.1016/J.OPHTHA.2021.04.027

Ting, D. S. W., Peng, L., Varadarajan, A. V., Keane, P. A., Burlina, P. M., Chiang, M. F., Schmetterer, L., Pasquale, L. R., Bressler, N. M., Webster, D. R., Abramoff, M., & Wong, T. Y. (2019). Deep learning in ophthalmology: The technical and clinical considerations. *Progress in Retinal and Eye Research*, *72*. https://doi.org/10.1016/J.PRETEYERES.2019.04.003

Tobin, K. W., Chaum, E., Priya Govindasamy, V., & Karnowski, T. P. (2007). Detection of anatomic structures in human retinal imagery. *IEEE Transactions on Medical Imaging*, *26*(12), 1729–1739. https://doi.org/10.1109/TMI.2007.902801

Tong, Y., Lu, W., Yu, Y., & Shen, Y. (2020). Application of machine learning in ophthalmic imaging modalities. *Eye and Vision 2020 7:1*, *7*(1), 1–15. https://doi.org/10.1186/S40662-020-00183-6

Tripathy, K., Sharma, Y., R, K., Chawla, R., Gogia, V., Singh, S., Venkatesh, P., & Vohra, R. (2015). Recent Advances in Management of Diabetic Macular Edema. *Current Diabetes Reviews*, *11*(2), 79–97. https://doi.org/10.2174/1573399811999150324120640

Triwijoyo, B. K., Sabarguna, B. S., Budiharto, W., & Abdurachman, E. (2020). Deep learning approach for classification of eye diseases based on color fundus images. *Diabetes and Fundus OCT*, 25–57. https://doi.org/10.1016/B978-0-12-817440-1.00002-4

Van der Velden, B. H. M., Kuijf, H. J., Gilhuijs, K. G. A., & Viergever, M. A. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*, *79*, 102470. https://doi.org/10.1016/J.MEDIA.2022.102470

Wan, S., Liang, Y., & Zhang, Y. (2018). Deep convolutional neural networks for diabetic retinopathy detection by image classification. *Computers & Electrical Engineering*, *72*, 274–282. https://doi.org/10.1016/J.COMPELECENG.2018.07.042

Wang, W., Yang, Y., Wang, X., Wang, W., & Li, J. (2019). Development of convolutional neural network and its application in image classification: a survey. *Https://Doi.Org/10.1117/1.OE.58.4.040901*, *58*(4), 040901. https://doi.org/10.1117/1.OE.58.4.040901

Wang, X., Lu, Y., Wang, Y., & Chen, W. B. (2018). Diabetic retinopathy stage classification using convolutional neural networks. *Proceedings - 2018 IEEE 19th International Conference on Information Reuse and Integration for Data Science, IRI 2018*, 465–471. https://doi.org/10.1109/IRI.2018.00074

Wang, X. N., Dai, L., Li, S. T., Kong, H. Y., Sheng, B., & Wu, Q. (2020). Automatic Grading System for Diabetic Retinopathy Diagnosis Using Deep Learning Artificial Intelligence Software. *Current Eye Research*, *45*(12), 1550–1555. https://doi.org/10.1080/02713683.2020.1764975

Weith, K., Hassan, T., Schmid, U., & Garbas Jeans-Uwe. (2018). Towards explaining deep learning networks to distinguish facial expressions of pain and emotions . *Comprehensible Machine Learning/Comprehensible AI*. https://www.researchgate.net/publication/329371877_Towards_explaining_deep_learning_networks_to_distinguish_facial_expressions_of_pain_and_emotions

Welikala, R. A., Dehmeshki, J., Hoppe, A., Tah, V., Mann, S., Williamson, T. H., & Barman, S. A. (2014).

Automated detection of proliferative diabetic retinopathy using a modified line operator and dual classification. *Computer Methods and Programs in Biomedicine*, *114*(3), 247–261. https://doi.org/10.1016/J.CMPB.2014.02.010

Wilkinson, C. P., Ferris, F. L., Klein, R. E., Lee, P. P., Agardh, C. D., Davis, M., Dills, D., Kampik, A., Pararajasegaram, R., Verdaguer, J. T., & Lum, F. (2003). Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*, *110*(9), 1677–1682. https://doi.org/10.1016/S0161-6420(03)00475-5

Willoughby, C. E., Ponzin, D., Ferrari, S., Lobo, A., Landau, K., & Omidi, Y. (2010). Anatomy and physiology of the human eye: effects of mucopolysaccharidoses disease on structure and function – a review. *Clinical & Experimental Ophthalmology*, *38*(SUPPL. 1), 2–11. https://doi.org/10.1111/J.1442-9071.2010.02363.X

World Health Organization. (2021, October). *Blindness and vision impairment*. https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment

Wu, L., Fernandez-Loaiza, P., Sauma, J., Hernandez-Bogantes, E., & Masis, M. (2013). Classification of diabetic retinopathy and diabetic macular edema. *World Journal of Diabetes*, *4*(6), 290. https://doi.org/10.4239/WJD.V4.I6.290

Xu, K., Feng, D., & Mi, H. (2017). Deep Convolutional Neural Network-Based Early Automated Detection of Diabetic Retinopathy Using Fundus Image. *Molecules (Basel, Switzerland)*, *22*(12). https://doi.org/10.3390/MOLECULES22122054

Xu, X., Lin, J., Tao, Y., & Wang, X. (2019). An improved densenet method based on transfer learning for fundus medical images. *Proceedings - 7th International Conference on Digital Home, ICDH 2018*, 137–140. https://doi.org/10.1109/ICDH.2018.00033

Zahoor, M. N., & Fraz, M. M. (2017). Fast Optic Disc Segmentation in Retina Using Polar Transform. *IEEE Access*, *5*, 12293–12300. https://doi.org/10.1109/ACCESS.2017.2723320

Zhang, J., Xie, B., Wu, X., Ram, R., & Liang, D. (2021). *Classification of Diabetic Retinopathy Severity in Fundus Images with DenseNet121 and ResNet50*. http://arxiv.org/abs/2108.08473

Zheng, J., Guo, L., Peng, L., Li, J., Yang, J., & Liang, Q. (2014). Fundus image based cataract classification. *IST 2014 - 2014 IEEE International Conference on Imaging Systems and Techniques, Proceedings*, 90–94. https://doi.org/10.1109/IST.2014.6958452

Zhu, X., Rangayyan, R. M., & Ells, A. L. (2011). Digital Image Processing for Ophthalmology: Detection of the Optic Nerve Head. *Synthesis Lectures on Biomedical Engineering*, *40*, 1–106. https://doi.org/10.2200/S00335ED1V01Y201102BME040