Universidad de Valladolid

Escuela Técnica Superior
de Ingenieros de Telecomunicación

Trabajo Fin de Máster

Máster en Ingeniería de Telecomunicación

Clasificación automática de la severidad de la Retinopatía Diabética mediante técnicas de Deep Learning

Autora:

Dª. María Herrero Tudela

Tutora:

Dra. Dª. María García Gadañón

Valladolid, 10 de Septiembre de 2021

Título:	**Clasificación automática de la severidad de la Retinopatía Diabética mediante técnicas de Deep Learning**
Autor:	**Dª. María Herrero Tudela**
Tutora:	**Dra. Dª. María García Gadañón**
Departamento:	**Departamento de Teoría de la Señal y Comunicaciones e Ingeniería Telemática**

Tribunal

Presidente:	**Dr. D. Roberto Hornero Sánchez**
Secretario:	**Dr. D. Carlos Gómez Peña**
Vocal:	**Dra. Dª. Miriam Antón Rodríguez**
Suplente:	**Dr. D. Miguel López-Coronado**
Suplente:	**Dr. D. Jesús Poza Crespo**
Suplente:	**Dr. D. Salvador Dueñas Carazo**

Fecha:	**10 de Septiembre de 2021**
Calificación:	

A mis padres

Agradecimientos

Me gustaría que estas líneas sirvieran para expresar mi más sincero agradecimiento a todas aquellas personas que con su ayuda han colaborado en la realización del presente trabajo.

En primer lugar, quiero dar las gracias a todos los miembros del Grupo de Ingeniería Biomédica de la Universidad de Valladolid. Quiero hacer mención especial a mis tutores, María y Rober. Muchas gracias por guiarme con infinita paciencia en la realización de este trabajo.

Tampoco puedo olvidar la inestimable ayuda que me han prestado muchas personas en el plano personal. A mis amigos, les agradezco su apoyo y paciencia. A mi familia, y en especial a mis padres, por dejarme volar sin condicionarme en nada, por creer en mí siempre y por su amor, que guía mi profesión y mi vida. Finalmente, me gustaría agradecer a Dani su apoyo, cariño y comprensión, tanto en los momentos buenos como en los momentos malos.

Muchas gracias a todos.

Resumen

La retinopatía diabética (RD) es una de las principales enfermedades crónicas discapacitantes y una de las principales causas de ceguera y discapacidad visual en los países desarrollados. Los estudios indican que el 90% de los casos pueden prevenirse mediante la detección precoz y el tratamiento adecuado. Los médicos utilizan el cribado ocular mediante imágenes de la retina para detectar las lesiones relacionadas con esta enfermedad en exámenes oftalmológicos periódicos. Debido al creciente número de personas diabéticas, la cantidad de imágenes que los oftalmólogos especialistas han de analizar manualmente se está volviendo inasequible. Además, la formación de nuevo personal para este tipo de diagnóstico basado en imágenes es larga, ya que requiere adquirir experiencia mediante la práctica diaria.

En este trabajo se propone el desarrollo y evaluación de un método automático para la clasificación del grado de severidad de la RD basado en el análisis de retinografías. Dicho método permitiría identificar las imágenes que muestren signos de RD y determinar qué imágenes pertenecen a pacientes que deban ser derivados a atención especializada de forma preferente u ordinaria, según el estadio de gravedad detectado por el sistema. Para este fin, se ha empleado como clasificador una red neuronal profunda combinada con diferentes técnicas de *deep learning* tales como *data augmentation*, *dropout*, *transfer learning* y *fine tuning*.

El método se evaluó en la base de datos pública de retinografías APTOS-2019, en tres escenarios diferentes: clasificación de la RD en cinco grados de severidad (sano, RD no proliferativa leve, RD no proliferativa moderada, RD no proliferativa severa y RD proliferativa), detección de la presencia de la RD (clasificación en sano vs RD) y clasificación en casos derivables vs no derivables. Estos escenarios se abordan habitualmente en los trabajos existentes sobre el diagnóstico de la RD. En el escenario de clasificación multiclase se ha obtenido un coeficiente kappa de 0.919, una especificidad de 96.31%, un *Area Under Curve* (AUC) de 0.93 y una precisión de 94.10%. Para el caso de la detección de la presencia de la RD se ha obtenido un coeficiente kappa de 0.978, una sensibilidad

de 97.99%, una especificidad de 100%, un AUC de 1 y una precisión de 98.91%. Finalmente, en la detección de los casos derivables y no derivables se ha conseguido un coeficiente kappa de 0.811, una sensibilidad de 85.15%, una especificidad del 100%, un AUC de 0.97 y una precisión de 90.71%. Además, como los resultados proporcionados por los modelos basados en *deep learning* son difíciles de interpretar, se ha incluido un análisis de los resultados utilizando técnicas de *Explainable Artificial Intelligence* (XAI). En concreto, se ha utilizado *SHapley Additive exPlanations* (SHAP).

Los resultados obtenidos permiten comprobar que es posible tanto la detección y clasificación de la RD en casos derivables y no derivables como la clasificación automática de la severidad de la RD mediante el análisis de retinografías de pacientes diabéticos. El método propuesto, por tanto, permitiría acortar el tiempo de obtención de un diagnóstico, reducir la carga de trabajo de los expertos oftalmólogos y, como consecuencia, los costes económicos asociados al tratamiento de la RD. Además, el modelo está diseñado de tal forma que se puede generalizar su posible aplicación a otros tipos de imágenes médicas y dominios de clasificación.

Palabras clave

Análisis de retinografías; clasificación automática; *deep learning*; redes neuronales convolucionales; retinopatía diabética.

Abstract

Diabetic retinopathy (DR) is a major chronic disabling disease and a leading cause of blindness and visual impairment in developed countries. Studies indicate that 90% of cases are preventable through early detection and appropriate treatment. Eye screening using retinal imaging is used by physicians to detect lesions related to DR. Due to the increasing number of diabetic patients, the amount of images that specialist ophthalmologists have to analyse manually is becoming unaffordable. Furthermore, the training of specialists for this type of image-based diagnosis is time-consuming, as it requires gaining experience through daily practice.

This work proposes the development and evaluation of an automatic method for the classification of DR severity based on the analysis of fundus images. This method would make it possible to identify images showing signs of DR and to determine which images belong to patients who should be referred to specialised care on a preferential or ordinary basis, depending on the severity detected by the system. For this purpose, a deep neural network was used as a classifier, in combination with different deep learning techniques such as data augmentation, dropout, transfer learning and fine tuning.

The method was evaluated on the APTOS-2019 public retinography database in three different scenarios: classification of DR into five severity grades (healthy, mild non-proliferative DR, moderate non-proliferative DR, severe non-proliferative DR and proliferative DR), detection of the presence of DR (classification into healthy vs. DR cases) and classification into referable vs. non-referable cases. These scenarios are commonly addressed in the existing literature related to the diagnosis of DR. In the multiclass classification scenario, a kappa coefficient of 0.919, a specificity of 96.31%, an Area Under Curve (AUC) of 0.93 and an accuracy of 94.10% were obtained. In the case of detecting the presence of DR, a kappa coefficient of 0.978, a sensitivity of 97.99%, a specificity of 100%, an AUC of 1 and an accuracy of 98.91% were obtained. Finally, in the detection of referable and non-referable cases, a kappa coefficient of 0.811, a sensitivity of 85.15%, a specificity of 100%, an AUC of 0.97 and an accuracy of 90.71% were

achieved. In addition, as the results provided by deep learning-based models are difficult to interpret, an analysis of the results using Explainable Artificial Intelligence (XAI) techniques has been included. Specifically, SHapley Additive exPlanations (SHAP) have been used.

The results obtained show that it is possible both to detect and classify DR in referable and non-referable cases and to automatically classify the severity of DR by analysing retinographs of diabetic patients. The proposed method would therefore shorten the time it takes to obtain a diagnosis, reduce the workload of ophthalmic experts and, as a consequence, the economic costs associated with the treatment of DR. Furthermore, the model is designed in such a way that it can be generalised to other types of medical imaging and classification domains.

Keywords

Automatic classification; convolutional neural networks; deep learning; diabetic retinopathy; retinal imaging.

Glosario de términos y acrónimos

Adam. *Adaptive moment estimation*

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

BD. Base de datos

CNN. Red neuronal convolucional (*Convolutional Neural Network*)

DE. Diabetes en el embarazo

DM. Diabetes Mellitus

DMG. Diabetes Mellitus Gestacional

ETDRS. Estudio para el tratamiento temprano de la Retinopatía Diabética (*Early Treatment Diabetic Retinopathy Study*)

GLCM. Matriz de coocurrencia de niveles de gris (*Gray Level Co-occurence Matrix*)

HOS. Espectros de alto orden (*Higher Order Spectra*)

IRMA. Anomalías microvasculares intrarretinianas (*Intraretinal Microvascular Abnormality*)

KL. Divergencia de Kullback-Liebler

KNN. Algoritmo de los k vecinos más cercanos (*K-Nearest Neighbors*)

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

OCTA. Angiografía OCT (*Optical Coherence Tomography Angiography*)

OMS. Organización Mundial de la Salud

RD. Retinopatía diabética

RDNP. Retinopatía diabética no proliferativa

ReLU. Unidad Lineal Rectificada (*Rectified Linear Unit*)

RGB. Espacio de color Rojo-Verde-Azul (*Red-Green-Blue*)

RLM. *Run Length Matrix*

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

SHAP. SHapley Additive exPlanations

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

SNE. Incrustación de vecinos estocásticos (*Stochastic Neighbor Embedding*)

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

t-SNE. Incrustación de vecinos estocásticos distribuidos en t (*t-distributed Stochastic Neighbor Embedding*)

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

Durante este capítulo inicial se presenta el contexto y la motivación principal detrás de este trabajo, los objetivos perseguidos y la estructura en la que se plasma toda esta información a lo largo del mismo.

El presente documento pretende mostrar todas las tareas de investigación realizadas para la realización del Trabajo Fin de Máster que permite la obtención del título de Máster Universitario en Ingeniería de Telecomunicación de la Universidad de Valladolid.

1.1. Bioingeniería e ingeniería biomédica

La ingeniería biológica o bioingeniería es una disciplina orientada a la investigación que busca aplicar los métodos de la ingeniería a la solución de los problemas de la biología y la medicina. La bioingeniería cubre diversos campos de investigación, como la biotecnología, los biomateriales, la ingeniería clínica o la ingeniería biomédica (Enderle & Bronzino, 2011; Mompín Poblet, 1988).

La ingeniería biomédica es una subdisciplina de la bioingeniería, en la que la ingeniería y la medicina intercambian conocimientos y metodología con el objetivo de controlar las enfermedades en las personas. La ingeniería biomédica surge de la necesidad de la aplicación de la ingeniería a los procesos y tareas de los especialistas en medicina. De esta manera, busca reducir la brecha tecnológica existente en el ámbito de la salud, facilitando a los especialistas las tareas que deben realizar, aumentando la productividad y mejorando el funcionamiento del sistema de salud. Los ingenieros biomédicos aplican principios de ingeniería eléctrica, química, óptica, mecánica y de otro tipo para comprender, modificar o controlar los sistemas biológicos. Por lo tanto, los ingenieros biomédicos son miembros del equipo de atención médica que buscan nuevas soluciones a los

problemas de salud que enfrenta la sociedad moderna (Bronzino, 2006; Enderle & Bronzino, 2011; Mompín Poblet, 1988).

1.2. Imagen médica. Retinografías.

Las imágenes médicas se emplean para obtener información acerca del cuerpo humano con el fin de diagnosticar, monitorizar o tratar afecciones médicas. Las técnicas y procesos utilizados para obtener imágenes de diversas partes del cuerpo humano permiten a los especialistas realizar diagnósticos cada vez más precisos y proporcionar tratamientos adecuados (Mikla & Mikla, 2014).

La automatización del análisis de imágenes médicas mediante la aplicación de métodos de procesado digital de imágenes y técnicas de reconocimiento de patrones, visión por ordenador, etc. permite la aceleración del proceso de diagnóstico (Toennies, 2017). Debido a la falta de expertos que puedan analizar las imágenes adquiridas, tarea que requiere mucho tiempo y esfuerzo, es necesario desarrollar técnicas de ayuda al diagnóstico de enfermedades (Toennies, 2017). Además de mejorar la eficiencia del diagnóstico y seguimiento de enfermedades, estas técnicas también permiten reducir la subjetividad de los observadores humanos (Toennies, 2017).

En la oftalmología, el estudio del fondo de ojo se ha convertido en una exploración rutinaria e indispensable para el diagnóstico y tratamiento de numerosos procesos patológicos oculares, como el diagnóstico de enfermedades que afectan a la retina (Suri *et ál.*, 2014).

La anatomía y fisiología ocular (Figura 1.1) es similar en la mayoría de los vertebrados. El ojo humano mide aproximadamente de 22 a 27 mm de diámetro anteroposterior. El globo ocular consta de tres capas primarias, siendo estas: (1) la capa de soporte más externa del ojo, que incluye la córnea y la esclerótica; (2) la capa media superior del ojo, que constituye la capa vascular central del globo, que abarca el iris, el cuerpo ciliar y la coroides; y (3) la capa interior del ojo, formada por la retina, el humor acuoso y el humor vítreo (Kels *et ál.*, 2015).

Centrándonos en la retina humana, esta constituye el tejido ocular más complejo y su estructura se encuentra altamente organizada (Riordan-Eva, Paul; Cunningham, 2012). La palabra retina procede del latín medieval *rete* o *retis*. Su

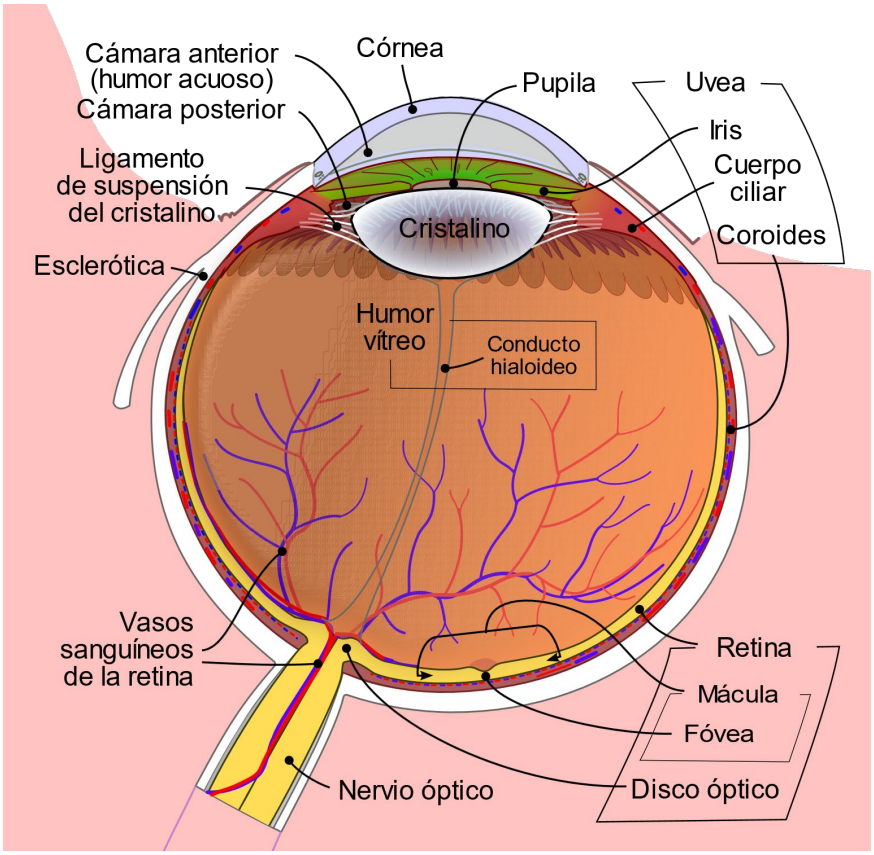


Figura 1.1 Corte del globo ocular en el que se señalan las estructuras visibles del ojo humano (Fuente: W. Commons, Schematic diagram of the human eye).

nombre se debe a la gran cantidad de vasos sanguíneos que la forman haciendo que su fina estructura se asemeje a la de una red (Real Academia Española, 2021).

A nivel macroscópico, la retina está formada por los elementos mostrados en la Figura 1.2 y que se detallan a continuación (Abràmoff *et ál.*, 2010):

- **Papila o disco óptico.** Es el punto de entrada del nervio óptico en el globo ocular. Debido a la ausencia de fotorreceptores también se conoce con el nombre de punto ciego. Bajo examen oftalmológico, se aprecia la papila como un disco circular o ligeramente ovalado con un diámetro aproximado de 1.5 mm y pigmento amarillo. A través del disco óptico entra al globo ocular la arteria central de la retina y sale la vena central de la retina. En el disco óptico encontramos también una excavación fisiológica conocida como cúpula o copa.

- **Arterias y venas.** Son las encargadas de proveer de oxígeno y nutrientes a la retina. La arteria central de la retina entra en el ojo a través del nervio óptico y se separa en dos ramas, que a su vez divergen formando una extensa red de capilares. Muchas de las enfermedades oculares afectan a estos vasos sanguíneos, bloqueándolas o haciéndolas más frágiles.
- **Mácula.** Región ovalada diferenciada del resto de la retina por su gran pigmentación y por la disminución del calibre de los vasos sanguíneos. Se encuentra en el centro de la retina y tiene un diámetro aproximado de 5 mm. Es la encargada tanto la visión central como de la visión en detalle y en movimiento.
- **Fóvea.** Se encuentra en el centro de la mácula y es el área de mayor agudeza visual. Es una hendidura con un diámetro aproximado de 1 mm que permite enfocar los rayos que llegan a la retina.
- **Retina periférica.** Es la encargada de la visión periférica, es decir, la de los rayos de luz que no están en nuestro foco central de visión.

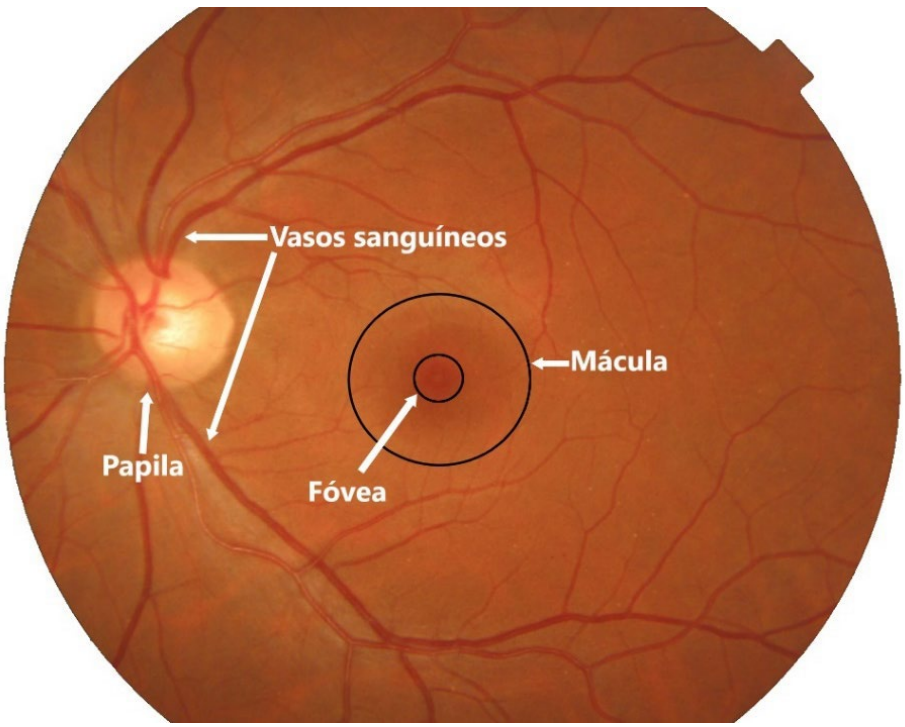


Figura 1.2 Imagen de retina en la que se marcan la papila, la mácula, la fóvea y los vasos sanguíneos.

A nivel microscópico, la retina es un tejido delgado, parcialmente transparente, compuesto por millones de células fotorreceptoras y neuronas (Savino & Danesh-Meyer, 2012). Existen dos tipos principales de fotorreceptores: los conos y los bastones (Figura 1.3). Estos receptores contienen unos compuestos químicos conocidos como fotopigmentos, los cuales tienen la propiedad de descomponerse ante la exposición a la luz, excitando en el proceso a las fibras nerviosas que salen del ojo (Hadjikhani & Tootell, 2000).

Los conos son células capaces de distinguir detalles finos y colores. En ellos encontramos tres tipos distintos de fotopigmentos que responden a longitudes de onda diferentes de la luz, lo que da lugar a los conocidos como colores primarios de la luz: rojo, azul, y verde. Los bastones, por el contrario, únicamente registran grises, y son más relevantes para la visión periférica y la vista en condiciones de baja iluminación (Savino & Danesh-Meyer, 2012).

Tanto la densidad como la distribución de los conos y bastones es diferente. Existe una alta concentración de conos en la fóvea, predominantemente sensibles al rojo y al verde, con una densidad mayor de 140.000 conos/mm^2. Los bastones tienen su máxima densidad en una zona situada a 20° de la fóvea, donde alcanzan una densidad de aproximadamente 160.000 bastones/mm^2. El número de conos y bastones disminuye rápidamente al alejarse del centro, es decir, la periferia casi no contiene conos ni bastones (Purves *et ál.*, 2001).

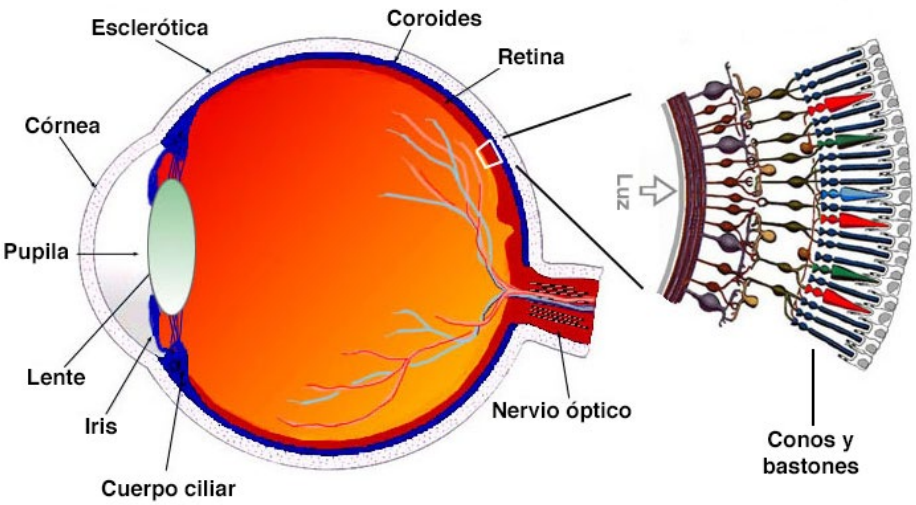


Figura 1.3 Corte del globo ocular donde se representa la situación de los conos y los bastones (Fuente: Blue Cone Monochromacy Families Foundation).

Ambos tipos de células están conectadas con las neuronas ganglionares a través de una capa de células bipolares. Las neuronas ganglionares transmiten las señales eléctricas a través del nervio óptico, para que finalmente sean procesadas por el córtex visual en el cerebro (Vilensky *et ál.*, 2015). Las células de la retina presentan grandes similitudes con las células del cerebro, apoyando la afirmación de que el sistema visual es una extensión del sistema nervioso central (Zhu *et ál.*, 2012).

Para permitir la formación de las imágenes en la retina, las estructuras del ojo humano deben ser transparentes. Esto permite que la anatomía del ojo pueda ser examinada de una manera no invasiva si se emplean las técnicas adecuadas (Abràmoff *et ál.*, 2010). Los oftalmólogos obtienen una visión directa de los vasos sanguíneos y los tejidos del sistema nervioso central (retina y nervio óptico), lo que les permite la identificación de efectos derivados de enfermedades infecciosas, autoinmunes y vasculares (Besenczi *et ál.*, 2016; Riordan-Eva, Paul; Cunningham, 2012). Entre las técnicas de obtención de imágenes de la retina cabe destacar las siguientes (García Gadañón, 2008; Martínez Rubio *et ál.*, 2012; Riordan-Eva, Paul; Cunningham, 2012; Stewart, 2017):

- **Oftalmoscopia**. Esta técnica permite observar el fondo de ojo sin obtener una imagen de él. Hay dos tipos de oftalmoscopios: *directo* e *indirecto*. El oftalmoscopio directo es un instrumento óptico que permite la observación del fondo de ojo mediante la proyección de luz sobre él a través de un prisma. La luz es reflejada en la retina y el especialista obtiene una imagen de la retina amplificada. Sin embargo, su inconveniente principal es la limitada capacidad para observar con claridad la zona más periférica de la retina. El oftalmoscopio indirecto es un instrumento que se coloca en la cabeza del observador, y permite la visualización de la retina mediante una potente fuente de luz que se hace coincidir con el eje de mirada. Con su mano, el examinador sostiene una lente a unos centímetros del ojo del paciente enfocando la luz en la retina, de tal manera que se forma una imagen de esta en el espacio entre el paciente y el examinador. Aunque esta técnica ofrece una visión más amplia de la retina, requiere de la dilatación de la pupila. No obstante, debido a su simplicidad, la oftalmoscopia indirecta se ha convertido en la técnica de examen de fondo de ojo más empleada en los servicios de oftalmología para realizar un primer examen clínico.

- **Angiografía fluoresceínica**. Se trata de una técnica aparecida a finales de los años 60 que requiere la inyección de fluoresceína en una vena del antebrazo del paciente. Posteriormente, es posible visualizar el fondo de ojo a través de determinados tipos de oftalmoscopios equipados con filtros especiales. La fluoresceína es un colorante cuyas moléculas emiten luz verde al ser estimuladas por luz azul. En las fotografías, este colorante resalta los detalles vasculares y anatómicos del fondo del ojo. La angiografía con fluoresceína es útil en el estudio de vasculopatías retinianas, anomalías del epitelio pigmentado y lesiones subyacentes.
- **Oftalmoscopia de láser de barrido** (*Scanning Laser Ophthalmoscope*, SLO). Se trata de una técnica aparecida en los años 80 que permite la obtención de imágenes de la retina con alta resolución y sin la necesidad de dilatar la pupila. Un pequeño punto de láser explora el fondo de ojo realizando barridos para iluminar los elementos sucesivos de la retina, a velocidades de hasta 24 milisegundos. La luz reflejada desde cada punto de la retina es capturada por un fotomultiplicador. La salida del fotomultiplicador se graba y se muestra en formato de vídeo digital. Además, el sistema es insensible a opacidades oculares que puedan causar diversos artefactos.
- **Tomografía de coherencia óptica** (*Optical Coherence Tomography*, OCT). Esta técnica no invasiva apareció en los años 90 y permite la visualización de cortes histológicos de la retina. La OCT del segmento posterior posibilita el análisis detallado de la papila, la capa de fibras nerviosas de la retina y la mácula. Para el análisis del segmento anterior se utiliza otro aparato de OCT, que proyecta un haz de luz infrarroja con una longitud de onda de 1300 nm. Este instrumento permite mediciones e imágenes de alta resolución de córnea, iris y otras estructuras intraoculares.
- **Angiografía por tomografía de coherencia óptica** (*Optical Coherence Tomography Angiography*, OCTA). Este procedimiento permite la observación de los vasos sanguíneos coroidales y retinianos sin el uso de un tinte inyectado. La OCTA puede detectar defectos de perfusión en los plexos capilares retinianos superficiales y profundos, y, podría reemplazar la angiografía con fluoresceína en la evaluación de la enfermedad vascular retiniana.

- **Autofluorescencia**. Es una técnica no invasiva que aprovecha la propiedad autofluorescente de la retina para la visualización de signos de daño oxidativo. Esta cualidad consiste en la emisión de una luz en el espectro de longitudes de onda de los 500 a los 700 nm. La autofluorescencia de la retina se debe, principalmente a un componente presente en el epitelio pigmentario llamado lipofucsina. La intensidad de la autofluorescencia es mayor cuanto mayor es la cantidad y distribución de la lipofuscina.

- **Retinografía**. Es una técnica no invasiva e indolora que consiste en la obtención de imágenes en color de la retina permitiendo la visión exacta de la papila, la mácula y los vasos sanguíneos. Las retinografías pueden ser de varios tipos, en función de la necesidad de usar colirios midriáticos o no, y del ángulo que se obtiene con la fotografía en una única captura. Por tanto, los retinógrafos pueden ser también de dos tipos: midriáticos y no midriáticos. La midriasis es el aumento del diámetro de la pupila mediante la aplicación de un colirio llamado tropicamida. La utilización de los retinógrafos no midriáticos tiene la ventaja de no ocasionar al paciente las molestias de la midriasis provocada, siendo estas la elevada sensibilidad a la luz y la visión borrosa. Esta técnica permite un registro permanente de las imágenes en soporte digital, haciendo posible su posterior análisis por parte de oftalmólogos o sistemas de diagnóstico asistido por ordenador.

1.3. Análisis de imágenes de fondo de ojo para diagnóstico asistido por ordenador

Las retinografías constituyen la modalidad de captura más utilizada para el *screening* de diversas enfermedades oculares. Esto se debe a que las imágenes de fondo de ojo se obtienen mediante un procedimiento no invasivo e indoloro, que puede ser realizado por técnicos y no exclusivamente por especialistas en oftalmología (Lin *et ál.*, 2002). Posteriormente, estas imágenes pueden transmitirse hacia instalaciones remotas en las que los expertos oftalmólogos realizan el correspondiente diagnóstico, permitiendo el diagnóstico de un mayor número de pacientes y la reducción de costes (Lin *et ál.*, 2002).

Sin embargo, el cuello de botella en cualquier campaña de *screening* se encuentra en el análisis de las imágenes. Con la creciente incidencia de

enfermedades oculares, el número de imágenes que han de ser revisadas por los expertos se ha incrementado notablemente. Esto, unido a la carencia de oftalmólogos especialistas, provoca que el tiempo para obtener una valoración clínica de una retinografía se incremente (Tozer *et ál.*, 2015).

Los métodos automáticos para detectar estas enfermedades suponen una importante herramienta para reducir estos inconvenientes y mitigar el coste en expertos de las campañas de *screening* (Michelson, 2015). Por medio de sistemas capaces de detectar lesiones o cambios patológicos en las imágenes, se puede generar un orden de los casos, identificando los más críticos, y permitiendo a los médicos decidir aquellos que deben analizarse primero (Michelson, 2015).

En general, los métodos automáticos para el estudio de enfermedades oftalmológicas pueden agruparse en las cuatro categorías que se muestran en la Tabla 1.1 (Muthu Rama Krishnan Mookiah *et ál.*, 2013). Los métodos basados en caracterización vascular correlacionan cambios en la distribución de los vasos sanguíneos de la retina con la existencia de alguna enfermedad. Por ejemplo, el calibre de los vasos y su dimensión fractal han sido identificados como potenciales indicadores de estadios tempranos de la retinopatía diabética (RD) (Cheung *et ál.*, 2009; Crosby-Nwaobi *et ál.*, 2012). Los enfoques basados en la detección de estructuras patológicas cuantifican la existencia de lesiones asociadas con las enfermedades, como las lesiones rojas, lesiones brillantes o neovascularizaciones (Roychowdhury *et ál.*, 2016; Seoud *et ál.*, 2015). Estos enfoques también incluyen otras estrategias basadas en el cálculo de descriptores globales a partir de imágenes de personas enfermas y sanas, y utilizan esos valores para discriminar los casos con patologías. Entre estos descriptores globales, se encuentran aquellos obtenidos utilizando espectros de alto orden (*Higher Order Spectra*, HOS) para caracterizar las imágenes, y una máquina de vectores de soporte (*Support Vector Machine*, SVM) que asigna a cada una de ellas un índice de riesgo de la

Propiedades vasculares	**Detección de estructuras patológicas**	**Caracterización por descriptores globales de imagen**	**Métodos de aprendizaje profundo**
Calibre, dimensión fractal	Lesiones rojas, lesiones brillantes, neovascularizaciones	HOS, Texturas (GLCMs y RLMs)	Redes neuronales convolucionales

Tabla 1.1 Clasificación de las herramientas para el diagnóstico y la detección asistida de enfermedades oculares.

enfermedad. También existen otros caracterizadores de texturas tales como las matrices de coocurrencia de niveles de gris (*Gray Level Co-occurence Matrices*, GLCMs) y las matrices de *run length* (*Run Length Matrices*, RLMs) (Acharya U *et ál.*, 2008; U Rajendra Acharya *et ál.*, 2012). Por último, los sistemas basados en *deep learning* utilizan redes neuronales convolucionales (*Convolutional Neural Networks*, CNNs) de manera que el sistema aprende de forma automática las características que permiten identificar estas enfermedades (Pratt *et ál.*, 2016). Es necesario tener en cuenta que esta clasificación propuesta no es estricta, ya que existen diversos algoritmos que ocupan más de una categoría en particular.

1.4. Diabetes mellitus

El término diabetes mellitus (DM) describe un trastorno metabólico caracterizado por hiperglucemia crónica con alteraciones del metabolismo de carbohidratos, grasas y proteínas como resultado de defectos en la secreción de insulina. La DM puede presentarse con síntomas característicos como sed, poliuria, visión borrosa y pérdida de peso. A menudo, los síntomas no son graves o pueden estar ausentes y, en consecuencia, la hiperglucemia suficiente para causar cambios patológicos y funcionales puede estar presente durante mucho tiempo antes del momento del diagnóstico. Los efectos a largo plazo de la DM incluyen enfermedades cardiovasculares, lesión de los nervios (neuropatía), enfermedad renal (nefropatía) y afecciones oculares, causantes de la RD, la pérdida de visión e incluso la ceguera. Sin embargo, si se logra un tratamiento apropiado de la DM, estas graves complicaciones se pueden retrasar o prevenir totalmente (Federación Internacional de Diabetes, 2019).

Actualmente, la DM afecta a más de 463 millones de personas en todo el mundo y se espera que su incidencia crezca hasta los 578,4 millones de personas para 2030. Asimismo, para 2045 esta cifra aumentaría a 700,2 millones (Federación Internacional de Diabetes, 2019).

Se han identificado los siguientes tres tipos principales de DM (Federación Internacional de Diabetes, 2019; Yanoff & Sassani, 2018):

- **Diabetes tipo I.** En esta forma de la enfermedad el sistema inmunitario del organismo ataca a las células β del páncreas que producen insulina. Como resultado, el organismo no produce insulina o la cantidad que produce no es suficiente. Este tipo de DM puede

aparecer a cualquier edad, aunque ocurre de forma más frecuente en niños y jóvenes. Para mantener el nivel de glucosa dentro de los valores apropiados las personas con esta enfermedad necesitan inyecciones de insulina, y es por ello por lo que también se llama DM insulinodependiente.

- **Diabetes tipo II.** La DM tipo II también es conocida como diabetes no insulinodependiente o diabetes de inicio en la edad adulta, y representa aproximadamente entre el 90% y el 95% de todos los casos totales de DM en el mundo. En este tipo de DM, la hiperglucemia resulta de la incapacidad de las células del cuerpo de responder de forma total a la insulina, lo que se conoce como "resistencia a la insulina". El resultado es un aumento de la producción de insulina que, con el tiempo, puede llegar a una producción de insulina inadecuada porque las células β pancreáticas no cumplen con la demanda del organismo. En general, esta enfermedad suele aparecer sin síntomas, siendo difícil determinar el momento exacto de su origen. Como consecuencia, el período prediagnóstico es, a menudo, largo, provocando que entre un tercio y la mitad de las personas con DM tipo II no reciban el diagnóstico correspondiente. Si no se identifica la enfermedad en sus estadios tempranos, en el momento del diagnóstico pueden estar presentes ciertas complicaciones como la RD, enfermedad que se ha convertido en una de las principales causas de ceguera a nivel mundial.

- **Diabetes mellitus gestacional.** Según la Organización Mundial de la Salud (OMS) y la Federación Internacional de Ginecología y Obstetricia, la hiperglucemia en el embarazo se clasifica como DM gestacional (DMG) o diabetes en el embarazo (DE). Por una parte, la DMG se diagnostica por primera vez durante el embarazo y puede ocurrir en cualquier momento de este período, siendo más frecuente después de la semana número 24. Por otra parte, la DE hace referencia a las mujeres embarazadas que previamente han sido diagnosticadas con diabetes o que padecen hiperglucemia diagnosticada por primera vez durante el embarazo, y que cumple con los criterios de la OMS sobre la diabetes durante el período de no embarazo. Cabe destacar que la DE puede aparecer en cualquier momento del periodo de embarazo, incluido el primer trimestre. Se calcula que la mayoría de

los casos de hiperglucemia que se producen en el embarazo (entre un 75% y un 90%) son DMG.

1.5. Retinopatía diabética

La RD es una consecuencia del daño microvascular producido por la DM (Abràmoff *et ál.,* 2010), y es una de las principales causas de ceguera a nivel mundial. Alrededor del 5% de las personas con DM tipo II padecen RD, y se espera que el número de pacientes afectados por esta enfermedad se vea significativamente incrementado en los próximos años (Abràmoff & Niemeijer, 2015).

Actualmente, no existe un método de prevención para la RD. No obstante, si se trata adecuadamente, es posible prevenir más del 90% de los casos de pérdida de visión. Antes de que la visión se vea afectada, aparecen cambios fácilmente detectables. Las lesiones típicas derivadas de la RD se pueden observar en la Figura 1.4 y se resumen en (Early Treatment Diabetic Retinopathy Study Research Group, 1991):

- **Exudados blandos**. También llamados manchas algodonosas. Se trata de engrosamientos isquémicos de la capa de fibras nerviosas que tapan los vasos retinianos. Presentan bordes difusos y un color blanquecino.
- **Exudados duros**. Depósitos lipídicos intrarretinianos de color amarillento brillante y bien definidos. Suelen encontrarse en la capa más externa de la retina.
- **Hemorragias.** Pequeñas manchas rojas con diversas formas y márgenes ligeramente definidos que aparecen como consecuencia de los puntos de sangrado en la retina. Poseen una forma redondeada cuando aparecen en zonas intermedias de la retina, o una forma más indefinida cuando se producen cerca de los vasos principales.
- **Microaneurismas**. Pequeñas protuberancias en los vasos sanguíneos de la retina. Se presentan como grupos de puntos rojos de color rojo oscuro con bordes muy definidos situados en zonas cercanas a las venas perimaculares o alrededor de las arterias temporales. Su número y extensión van creciendo con la progresión de la enfermedad

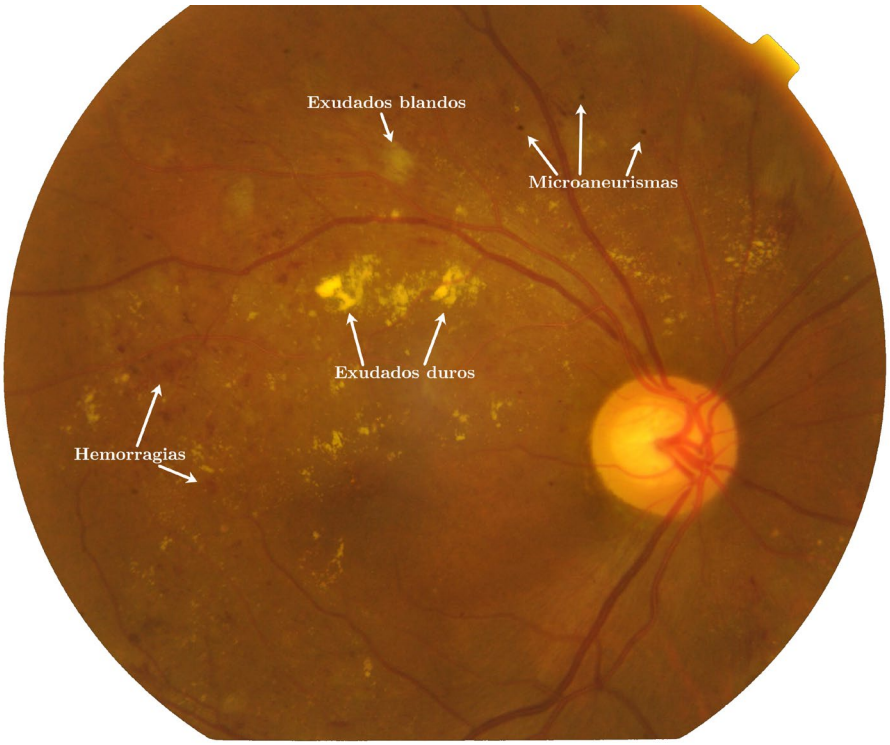


Figura 1.4 Retinografía de un sujeto con RD, donde aparecen microaneurismas, hemorragias y exudados.

Teniendo en cuenta las lesiones anteriormente mencionadas, existen distintos grados de severidad de la RD, por lo que es necesaria una clasificación de la misma. Debido a esto, el grupo conocido como *Global Diabetic Retinopathy Project Group*, elaboró una clasificación publicada en 2003 por la Academia Americana de Oftalmología que se ha convertido en el estándar de trabajo para una práctica clínica de calidad (Wilkinson *et ál.*, 2003). Esta clasificación es la que se emplea en el protocolo ETDRS (*Early Treatment Diabetic Retinopathy Study*). Este protocolo, aceptado como estándar para la detección y graduación de la RD utiliza 7 fotos estándar de pacientes con midriasis y evaluadas por personal cualificado (Early Treatment Diabetic Retinopathy Study Research Group, 1991).

La severidad de la RD se divide en 2 grandes niveles, asociándose las primeras fases a la RD no proliferativa (RDNP), y las más avanzadas a la RD proliferativa. Con más detalle, se distinguen 5 niveles definidos por la observación oftalmoscópica del fondo de ojo y que se resumen en la Tabla 1.2 (Aliseda & Berástegui, 2008; Klein *et ál.*, 2009):

Grado 0. Sin retinopatía aparente

Los exámenes realizados a las imágenes de fondo de ojo de esta clase no muestran ninguna anomalía, ni en forma de microaneurismas ni en formas más complejas. Un paciente diabético sin RD tiene una probabilidad inferior al 1% de desarrollar una RD proliferativa en los siguientes cuatro años.

Grado 1. RDNP leve

Se trata de la etapa más temprana de la enfermedad. En esta etapa, los microaneurismas son la única anomalía que se encuentra en los exámenes. Los pacientes diabéticos de grado 1 tienen una probabilidad inferior al 5% de desarrollar una RD proliferativa en los siguientes cuatro años.

Grado 2. RDNP moderada

En esta etapa, los vasos sanguíneos en las retinas se hinchan. Se observa la aparición de algunas hemorragias retinianas o microaneurismas, pudiendo existir exudados duros o lipídicos y blandos o algodonosos, todo ello acompañado de dilataciones venosas.

Grado 3. RDNP severa

En esta etapa pueden presentarse cualquiera de las siguientes alteraciones: hemorragias intrarretinianas severas, dilataciones venosas o anomalías microvasculares intrarretinianas (*Intraretinal Microvascular Abnormality,* IRMA). Los IRMA son vasos finos tortuosos anormales anexos a la arquitectura vascular retiniana. Estas alteraciones provocan que los vasos sanguíneos se bloqueen haciendo que llegue menos sangre a ciertas áreas de la retina. La falta de sangre estimula que se envíe una señal a estas áreas retinianas para crear nuevos vasos sanguíneos formando tejido cicatricial.

Los pacientes con RDNP severa tienen un 17% de posibilidades de desarrollar RD proliferativa en los siguientes tres años.

Grado 4. RD proliferativa

Es la etapa más avanzada de la enfermedad. En ella, los vasos sanguíneos nuevos, frágiles y anormales crecen en la retina o el nervio óptico. Los exámenes oculares detectan una neovascularización definida (aparición de nuevos vasos sanguíneos en la retina) o hemorragias prerretinianas o vítreas.

Nivel de gravedad	Observaciones
Grado 0: No RD aparente	Sin ninguna anomalía
Grado 1: RDNP leve	Presencia de algunos microaneurismas
Grado 2: RDNP moderada	Presencia de más microaneurismas pero menos que en el grado 3
Grado 3: RDNP severa	Alguno de los siguientes: - Más de 20 hemorragias intrarretinales - Dilataciones venosas - Anomalías microvasculares en la retina - Ningún signo de RDP
Grado 4: RD proliferativa	Alguno de los siguientes: - Neovascularización - Hemorragia vítrea

Tabla 1.2 Niveles de gravedad de la Retinopatía Diabética en función de las lesiones observadas (Wilkinson *et ál.*, 2003).

1.6. Hipótesis de trabajo

Según el informe mundial sobre la visión realizado por la OMS, al menos 2200 millones de personas padecen algún tipo de discapacidad visual. Entre ellas, cerca de 1000 millones tienen una deficiencia visual que podría haberse evitado o que aún no se ha tratado (Organización Mundial de la Salud, 2020). El informe detalla 7 principales causas de discapacidad visual entre las que se encuentra la RD**.**

La RD progresa en cuatro etapas. Cada etapa tiene sus características particulares, por lo que, si alguna de ellas no es tenida en cuenta es posible que el diagnóstico sea incorrecto. Al menos el 56% de los casos nuevos de esta enfermedad podrían reducirse con un tratamiento y control ocular adecuados (Rohan *et ál.*, 1989). Sin embargo, la detección temprana es difícil ya que la enfermedad es asintomática hasta sus estadios avanzados.

Además, como se ha visto en la literatura (Krause *et ál.*, 2018) los especialistas no siempre coinciden en la valoración del diagnóstico, como se muestra en la Figura 1.5. En ausencia de enfermedad y cuando las lesiones son evidentes, los médicos suelen estar de acuerdo. Sin embargo, en estadios intermedios la valoración difiere (Krause *et ál.*, 2018).

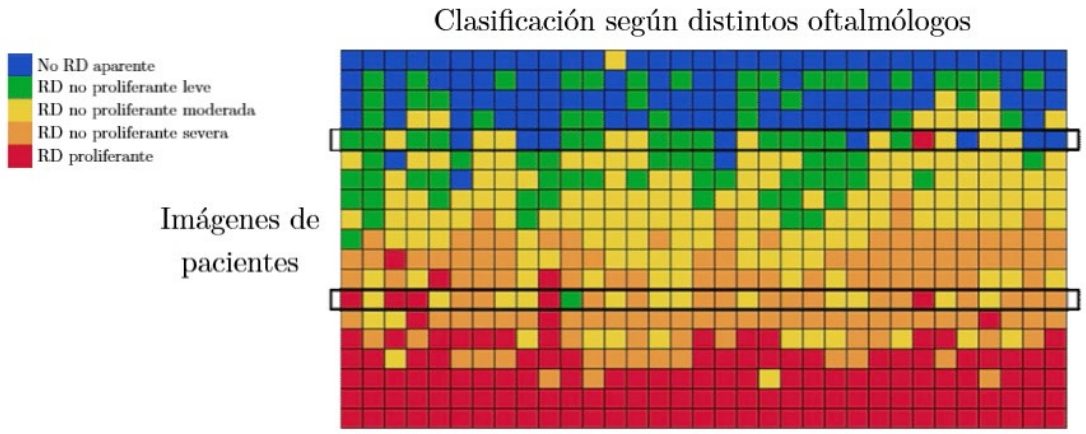


Figura 1.5 Diferencia en el diagnóstico de la RD realizado por distintos oftalmólogos para la misma imagen de fondo de ojo (adaptado de Krause *et ál.*, (2018)).

Con la creciente incidencia de la diabetes, el número de imágenes que han de ser revisadas por los expertos ha sufrido también un incremento importante. Además, alrededor del 90% de las imágenes revisadas en un programa de cribado no presentan signos de RD (Abramoff & Suttorp-Schulten, 2005). En este sentido, se plantea la hipótesis de que una herramienta de clasificación automática de la severidad de la RD podría acelerar el proceso de diagnóstico de dicha enfermedad, permitiendo el examen de un elevado número de imágenes en menor tiempo. Asimismo, permitiría seleccionar únicamente aquellas imágenes que requieren un examen adicional por parte del oftalmólogo, reduciendo de forma significativa su carga de trabajo.

En este trabajo se pretende diseñar un sistema automático de clasificación de la RD en sus diferentes etapas. Para ello, se han estudiado y analizado diversos tipos de arquitecturas de CNNs con el objetivo de obtener un método preciso y efectivo capaz de funcionar con imágenes muy variadas.

1.7. Objetivos del TFM

El objetivo principal de esta investigación ha sido el desarrollo de un sistema de clasificación automática de la severidad de la RD basado en *deep learning*. Para conseguir este objetivo, se han establecido una serie de objetivos más específicos que se detallan a continuación:

1. Estudio de la bibliografía disponible relacionada con el análisis de retinografías, comprensión de los conceptos necesarios para la realización de este trabajo e identificación de las técnicas empleadas en estudios previos. Estos conceptos abarcan desde la definición de la RD hasta la selección de los métodos de clasificación de imágenes que se aplicaron para realizar el trabajo. De esta revisión surgen las bases sobre las que se fundamenta el algoritmo desarrollado.

2. Revisión de las bases de datos públicas de retinografías con el objetivo de encontrar una que incluya imágenes de características variables. Además, será necesario que incluya imágenes de pacientes con distintos grados de severidad de la RD en número suficiente para los algoritmos de *deep learning* considerados en este TFM.

3. Diseño, desarrollo y evaluación de diversas arquitecturas de CNNs para la clasificación de la severidad de la RD a través del lenguaje de programación Python, y más concretamente, mediante las librerías *Keras* y *TensorFlow*.

4. Comprobación del funcionamiento del método desarrollado en la BD de retinografías y obtención de los resultados del estudio.

5. Análisis de los resultados obtenidos para verificar la idoneidad del método en la clasificación de los grados de severidad de la RD.

6. Extracción de conclusiones a partir del análisis de los resultados obtenidos.

1.8. Metodología empleada

Para la consecución de los objetivos perseguidos en este TFM, se ha seguido la metodología de investigación mostrada en la Figura 1.6. En este diagrama de bloques se representan las diferentes etapas del trabajo, que se detallan a continuación.

1. Búsqueda de la información necesaria para comprender el problema a resolver. Esto incluye conocer la información más relevante relativa a la DM y a la RD, así como de los métodos de clasificación automática de la RD basados en CNNs desarrollados en trabajos previos.

2. Familiarización con el lenguaje de programación Python y, en especial, con las librerías *Keras* y *TensorFlow* utilizadas en el contexto de *deep learning*.

3. Implementación del método seleccionado. Para llevar a cabo esta tarea, se utilizaron las funcionalidades del lenguaje de programación Python.

4. Procesado de las imágenes y obtención de resultados. Para la obtención de resultados, en primer lugar, se han buscado los parámetros óptimos del método empleando un subconjunto de las imágenes de la BD, lo que se conoce como conjunto de entrenamiento. Una vez determinada la configuración más adecuada de acuerdo con los datos de entrenamiento, se obtienen los resultados sobre un conjunto diferente de imágenes, no empleadas antes, lo que se conoce como conjunto de test.

5. Análisis de resultados y extracción de conclusiones. Documentación del trabajo desarrollado.

1.9. Estructura del documento

En este apartado se describe la organización de esta memoria. En este primer capítulo de introducción se ha presentado el problema a resolver en este TFM, así como los objetivos del mismo y la metodología seguida. Se complementa con los capítulos que se describen a continuación:

- En el capítulo 2 "Revisión del estado de la técnica" se presenta una revisión del estado de la técnica en la tarea de clasificación automática de la severidad de la RD. Se hace especial hincapié en aquellos trabajos que emplean CNNs para la clasificación de las imágenes, con el objetivo de situar este trabajo en el contexto adecuado.

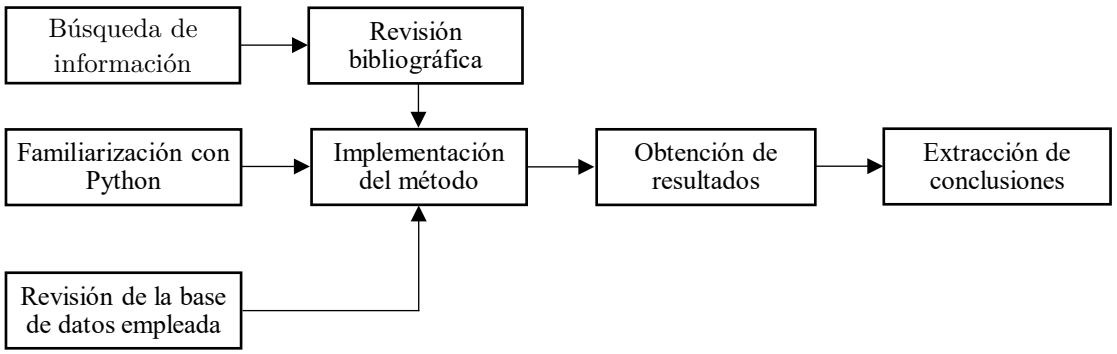


Figura 1.6 Esquema de la metodología seguida en este trabajo.

- En el capítulo 3 “Materiales y métodos” se presenta la BD utilizada en este TFM, así como el método aplicado.
- En el capítulo 4 “Resultados” se presentan los resultados más significativos del método desarrollado sobre la BD empleada.
- En el capítulo 5 “Discusión” se interpretan los resultados obtenidos en este TFM. Además, se comparan estos resultados con los obtenidos en otros estudios previos.
- En el capítulo 6 “Conclusiones y líneas futuras” se recogen las conclusiones más relevantes extraídas de este estudio. Asimismo, se detallan las principales aportaciones del trabajo desarrollado y se indican las posibles líneas de trabajo futuro.

Capítulo 2

Revisión del estado de la técnica

En la actualidad existe un creciente interés por el desarrollo de sistemas y algoritmos que permitan detectar la presencia de determinadas enfermedades oculares en un elevado número de sujetos. Entre ellos, el cribado de la población diabética con el objetivo de determinar aquellos pacientes que han desarrollado RD es una práctica cada vez más frecuente que reúne los esfuerzos de optometristas, médicos especialistas en diabetes y otros profesionales sanitarios (Barry *et ál.*, 2006).

En este capítulo se incluye una revisión de los principales trabajos que se han realizado para la detección y clasificación de la RD a partir de imágenes de fondo de ojo. Hoy en día, casi la totalidad de los nuevos modelos publicados en este campo son modelos basados en *deep learning*. Sin embargo, se comenzará analizando los modelos basados en *machine learning* que precedieron a los actuales.

2.1. Métodos basados en Machine Learning

Los modelos que se basan en *machine learning* para la detección de patologías en imágenes de fondo de ojo requieren, en cada imagen, de la extracción de una gran cantidad de características, de forma manual por los investigadores. Además, para la extracción de estas características es necesario conocimiento experto en la materia.

Estas técnicas se basan en detectar en las imágenes las lesiones que caracterizan la RD. Como se ha comentado anteriormente, estas lesiones son: exudados, microaneurismas y hemorragias. En el caso de la RD proliferativa, es posible encontrar también neovascularización. A partir de características obtenidas mediante técnicas de procesamiento digital de imagen y el uso de

clasificadores basados en *machine learning* es posible la detección de la enfermedad, e incluso, la estimación de su gravedad (Gulshan *et ál.*, 2016).

Muchos modelos se basan en la longitud, tamaño o posición de los vasos sanguíneos para determinar la existencia o no de enfermedad. Por ejemplo, Acharya *et ál.* (2009) aplicaron determinadas técnicas al canal verde de las imágenes de fondo de ojo, consiguiendo aislar los vasos sanguíneos del resto de la imagen. Otros autores (Tolias & Panas, 1998; Vlachos, 2009) se basan en la detección y seguimiento de las líneas centrales de los vasos sanguíneos. También encontramos estudios basados en la utilización de filtros adaptados de dos dimensiones (Luo *et ál.*, 2002; Mookiah *et ál.*, 2013). A partir de técnicas de preprocesado, existen sistemas capaces de detectar anchuras anormales en los vasos sanguíneos, indicio de la existencia de RD.

La presencia de exudados es una de las señales más características de la RD. Para su detección normalmente es necesario eliminar los vasos sanguíneos y el disco óptico de las imágenes. Tras esto, es posible detectar los exudados mediante una secuencia de algoritmos de procesamiento de imagen (U R Acharya *et ál.*, 2009). También es posible la detección de la RD a partir de la presencia de microaneurismas, donde Jelinek *et ál.* (2006) han conseguido una sensibilidad del 85% y una especificidad del 90%. Por otra parte, Quellec *et ál.* (2008) demostraron que el uso de la transformada *Wavelet* también resulta eficaz en la detección de microaneurismas.

Aunque las técnicas explicadas hasta el momento basan su predicción en la detección individual de alguna de las lesiones típicas de la RD, existen métodos que son capaces de detectar simultáneamente los tres tipos de lesiones y realizar predicciones en base a la presencia o no de cada lesión. Este es el caso del trabajo de Ege *et ál.* (2000), en el que se analizaron varios clasificadores estadísticos: clasificador bayesiano, clasificador basado en la distancia de Mahalanobis y el algoritmo de los k vecinos más cercanos (*K-Nearest Neighbors*, KNN). El clasificador de Mahalanobis presentó los mejores resultados: se detectaron microaneurismas, hemorragias, exudados y manchas algodonosas con una sensibilidad de 69, 83, 99 y 80%, respectivamente. Reza & Eswaran (2011) diseñaron un sistema que permitía detectar las lesiones brillantes con una precisión media del 97%.

Las investigaciones mencionadas anteriormente trataban de predecir una variable binaria que indicaba la presencia o no de RD basándose en las lesiones

de las imágenes. Sin embargo, otros estudios han tratado de clasificar el tipo de RD en proliferativa o no proliferativa. Mediante el uso de redes neuronales, Mookiah *et ál.* (2013) consiguieron una sensibilidad y especificidad de más del 96%. Acharya *et ál.* (2009) emplearon una SVM consiguiendo una sensibilidad de más del 82% y una especificidad del 86% en la detección de los 5 grados de la RD. En la misma línea, en trabajos posteriores se logró obtener una precisión de más del 90% en la detección de todos los grados de la RD (U Rajendra Acharya *et ál.*, 2012).

En la Tabla 2.1 se muestra un resumen de los diferentes métodos basados en *machine learning* descritos para la detección y clasificación de la RD.

Autor/res, año	**Descripción breve del método**	**Resultados sobre el conjunto de test**			
		AUC	**Precisión**	**Sensibilidad**	**Especificidad**
Acharya *et ál.* (2009)	Aislamiento de los vasos sanguíneos mediante SVM	-	-	82%	86%
Vlachos (2009)	Seguimiento de las líneas centrales de los vasos sanguíneos	-	92.9%	74.7%	95.5%
Jelinek *et ál.* (2006)	Detección de microaneurismas	-	-	85%	90%
Quellec *et ál.* (2008)	Transformada Wavelet	-	-	89.62%	89.50%
Ege *et ál.* (2000)	Clasificador bayesiano, clasificador basado en la distancia de Mahalanobis y KNN	-	-	99%	-
Reza & Eswaran (2011)	Detección de lesiones brillantes	-	97%	-	-
Mookiah *et ál.*, (2013)	Redes neuronales	-	-	96%	96%
U Rajendra Acharya *et ál.* (2012)	Extracción de características y SVM	0.895	100%	90%	-

Tabla 2.1 Comparación de métodos basados en *machine learning* para la detección de la RD.

2.2. Métodos basados en Deep Learning

Las aproximaciones que se basan en *deep learning*, haciendo uso de CNNs, representan los métodos más actuales en la mayor parte de tareas de análisis de imágenes médicas. Una de las principales ventajas de este tipo de técnicas es que se elimina la necesidad de la extracción manual de características en base a un conocimiento experto. Esto significa que la red se alimenta directamente de las imágenes y se realiza automáticamente el análisis y la extracción de características que permitan distinguir las distintas clases del problema.

En este contexto, existen dos principales grupos de algoritmos de detección de RD basados en *deep learning*: aquellos que tratan de detectar y localizar las lesiones características de la RD, y aquellos que detectan y clasifican directamente la enfermedad, sin enfocarse en la localización de lesiones concretas (Crosby-Nwaobi *et ál.*, 2012; Gargeya & Leng, 2017). De ambos grupos de algoritmos, el que se va a analizar es el segundo, puesto que es la metodología que se va a emplear en este trabajo. La gran ventaja de estos modelos es que no es necesario poseer las anotaciones de la localización de las lesiones de las imágenes para entrenar la red. Únicamente se debe disponer de un conjunto de imágenes con las etiquetas de cada clase: ausencia de enfermedad, RD no proliferativa leve, RD no proliferativa moderada, RD no proliferativa severa y RD proliferativa. La red CNN aprende de manera automática las características de las imágenes y las clasifica (Zhao *et ál.*, 2018). Sin embargo, su aplicabilidad a problemas de análisis de imágenes médicas está limitada por la carencia de grandes conjuntos de datos anotados, necesarios para el entrenamiento de estas redes (Zhang *et ál.*, 2019).

Para solventar este problema, diversos autores han propuesto pre-entrenar las redes neuronales con datos pertenecientes a otro contexto distinto del dominio de aplicación de interés, y aplicarlas posteriormente a la tarea que se pretende resolver. Este enfoque consiste en el entrenamiento de las redes utilizando grandes volúmenes de imágenes anotadas correspondientes a un dominio diferente, y posteriormente transferirlas al problema en cuestión *(Bar et ál.*, 2015; Ciompi *et ál.*, 2015; Pan & Yang, 2010).

Esta transferencia mencionada se basa en alimentar la red con las nuevas imágenes y extraer los vectores de características que la CNN calcula. Dichos vectores sirven para entrenar un clasificador que está exclusivamente dedicado a resolver el problema de clasificación deseado. Este enfoque parte de la hipótesis

de que, si los datos empleados en la fase de entrenamiento son lo suficientemente grandes y diversos, la red aprenderá a distinguir características útiles en diferentes contextos de clasificación de imágenes (Shin *et ál.*, 2016).

En el contexto de la detección y clasificación de la RD, casi la totalidad de los modelos de este tipo han empleado CNNs pre-entrenadas. Por ejemplo, Gulshan *et ál.* (2016) utilizaron una CNN profunda (*Deep Convolutional Neural Netowork*, DCNN) basada en la arquitectura Inception para para desarrollar un sistema de clasificación binaria de la RD, en la que se distinguía la presencia o ausencia de RD. Otro ejemplo es el estudio de Gargeya & Leng (2017) en el que desarrollaron un sistema de ayuda a la detección de la RD. Este sistema también se basaba en una DCNN que procesaba a la entrada retinografías en color, y las clasificaba como sanas o con RD. Otros métodos de *deep learning* detectaban la presencia de RD mediante el análisis de imágenes de fondo de ojo centradas en la mácula y el disco óptico como es el caso de Abràmoff *et ál.* (2018). Con el mismo procedimiento, Ting *et ál.* (2017) emplearon la arquitectura VGG-19 y Li et ál. (2018) utilizaron la arquitectura Inception-V3.

Sin embargo, otros investigadores han tratado de clasificar los diferentes niveles de severidad de la RD. Krause *et ál.* (2018) utilizaron una red Inception-V3 para clasificar la RD en 5 niveles: sin RD, RDNP leve, RDNP moderada, RDNP severa y RD proliferativa. Otros autores (Colas *et ál.*, 2016) entrenaron una CNN con 70000 retinografías con diferentes grados de RD. Su sistema logró un área bajo la curva característica operativa del receptor de 0.946. En la misma línea, Gwenolé Quellec *et ál.* (2017) diseñaron un modelo que conseguía un área bajo la curva característica operativa del receptor de 0.954 entrenando una CNN con más de 110000 imágenes de fondo de ojo. El método de Costa & Campilho (2017) logró un área bajo la curva de 0.97 utilizando el algoritmo *Bag-of-visual-words* junto con redes neuronales. Otros estudios entrenaron una CNN con más de 50000 imágenes de distintos grados de RD, logrando un valor de índice kappa de 0.925 en este problema de clasificación (Tymchenko *et ál.*, 2020). Por último, Galdran *et ál.* (2019) también hacen uso de CNNs y una BD de más de 40000 retinografías para obtener un sistema con más de un 90% de precisión en la clasificación de 4 clases de la RD.

En la Tabla 2.2 se muestra un resumen de los diferentes métodos basados en *deep learning* descritos para la detección y clasificación de la RD.

Autor/res, año	**Descripción breve del método**	**Resultados sobre el conjunto de test**				
		Coeficiente kappa	**AUC**	**Precisión**	**Sensibilidad**	**Especificidad**
Colas *et ál.* (2016)	CNN	-	0.946 (Kaggle)	-	96.20% (Kaggle)	66.6% (Kaggle)
Gulshan *et al.* (2016)	DCNN con arquitectura Inception-V3	-	0.94 (Messidor-2)	-	96.10% (Messidor-2)	93.90% (Messidor-2)
Gargeya & Leng (2017)	DCNN	-	0.94 (Messidor-2)	-	93.00% (Messidor-2)	87.00% (Messidor-2)
Gwenolé Quellec *et ál.* (2017)	CNN	-	0.954 (Kaggle) 0.949 (e-ophtha)	-	-	-
Ting et ál. (2017)	2 CNN con arquitectura VGG-19	-	0.936 (SiDRP 2014-2015)	-	90.5% (SiDRP 2014- 2015)	91.6% (SiDRP 2014- 2015)
Abramoff *et ál.* (2018)	CNN con imágenes centradas en el DO y en la mácula	-	-	-	80.70% (FDA Pivotal Trial)	89.80% (FDA Pivotal Trial)
Costa & Campilho (2017)	*Bag-of-visual-words*	-	0.97 (Messidor-2)	-	-	-
Krause *et ál.* (2018)	CNN con arquitectura Inception-V3	-	0.986 (EyePACS-2)	-	97.10% (EyePACS-2)	92.3% (EyePACS-2)
Li et al. (2018)	CNN con arquitectura Inception-V3	-	0.955 (NIEHS)	97.1% (NIEHS)	92.5% (NIEHS)	98.5% (NIEHS)
Galdran *et ál.* (2019)	CNN	-	0.70-0.96 (Diagnos)	-	-	-
Tymchenko *et ál.* (2020)	CNN	0.925 (APTOS 2019)	-	-	99%	99%

Tabla 2.2 Comparación de métodos basados en *deep learning* para la detección de la RD.

Capítulo 3

Materiales y métodos

El análisis automático de imágenes de fondo de ojo para la clasificación de la RD puede resultar muy útil en la ayuda al diagnóstico de esta enfermedad. Como se ha mostrado en el capítulo 2 de este TFM, se han desarrollado multitud de técnicas para la detección y clasificación de la RD a partir de imágenes de fondo de ojo. Entre ellas, resultan especialmente relevantes para este estudio aquellas que emplean CNNs.

En este capítulo se presentan las imágenes empleadas en este trabajo, junto con la descripción de las características y métodos empleados en el desarrollo del algoritmo propuesto en la tarea de clasificación automática de la RD en imágenes retinianas. En este trabajo, las CNNs se emplean como parte principal del algoritmo de clasificación de la severidad de la RD.

3.1. Base de datos de retinografías

La BD empleada en este trabajo pertenece a la Sociedad de Teleoftalmología de Asia Pacífico (*Asia Pacific Tele-Ophthalmology Society*, APTOS), que tiene como objetivo reunir a médicos, investigadores, técnicos, institutos y organizaciones para formar una alianza que promueva la comunicación, intercambio y colaboración en teleoftalmología. APTOS proporciona una plataforma en la que los profesionales de la atención oftalmológica pueden compartir conocimientos y colaborar para brindar una atención oftalmológica universal eficiente, accesible y de calidad en toda la región (Asia Pacific Tele-Ophthalmology Society, 2021).

El conjunto de datos utilizado para esta investigación pertenece a la BD pública *Kaggle APTOS 2019 Blindness Detection* (APTOS-2019). Dicha BD está formada por un total de 3662 imágenes de retina recogidas en múltiples clínicas

bajo una variedad de condiciones de imagen. Las imágenes de fondo de ojo pertenecientes a este conjunto de datos se clasifican en cinco clases:

- Ausencia de RD (clase 0).
- RDNP leve (clase 1).
- RDNP moderado (clase 2).
- RDNP severa (clase 3).
- RD proliferativa (clase 4).

Sin embargo, como se puede observar en los datos de la Tabla 3.1, el gran problema del conjunto de imágenes utilizado, que ha condicionado en gran medida la forma de trabajar con él, es el gran desbalanceo existente entre las clases. La clase predominante, las imágenes de retinas sanas, contiene el 50% del total de imágenes. Por el contrario, la clase minoritaria, las imágenes de retinas con RDNP severa, únicamente contiene 295 imágenes, aproximadamente el 5% del total.

La resolución de las imágenes que componen la BD es variada. Encontramos que la mayoría de ellas presentan una resolución de 1050x1050 píxeles. Sin embargo, también hay imágenes con resoluciones superiores, de 2416x1736 píxeles, y con resoluciones inferiores de 819x614 píxeles, entre otras. En la Tabla 3.2 se muestra un resumen con la cantidad de imágenes de cada resolución. El formato de todas ellas es PNG con una profundidad de 24 bits. Esto implica que cada píxel podrá tomar 256 valores de intensidad diferentes en el rango 0 – 255, para cada uno de los tres canales de color RGB.

Clase	**Total de imágenes**	**% de la BD completa**
Grado 0: No RD aparente	1805	50%
Grado 1: RDNP leve	370	10%
Grado 2: RDNP moderada	999	27%
Grado 3: RDNP severa	193	5%
Grado 4: RD proliferativa	295	8%

Tabla 3.1 Cantidad de imágenes de cada tipo en el conjunto de datos utilizado.

Dimensiones	Total de imágenes
474x358	2
640x480	42
819x614	287
1050x1050	974
1467x1110	2
1476x1117	14
1504x1000	92
1844x1226	61
2048x1536	351
2144x1424	28
2146x1764	1
2416x1736	638
2588x1958	533
2896x1944	34
3216x2136	410
3388x2588	141
4288x2848	52

Tabla 3.2 Cantidad de imágenes de cada resolución en el conjunto de datos utilizado.

En este TFM, las imágenes se han dividido de manera aleatoria para formar un conjunto de entrenamiento, uno de validación y otro de test. Para la división de la BD se ha elegido una proporción 80:10:10. Esto da como resultado un conjunto de entrenamiento formado por 2930 imágenes (80% de 3662) mientras que los conjuntos de validación y test están formados por 366 imágenes (10% de 3662).

3.2. Preprocesado de retinografías

El entrenamiento, la validación y el test del modelo se realizaron con versiones preprocesadas de las imágenes originales. En primer lugar, se redimensionaron todas las imágenes a una resolución de 512 x 512 píxeles con el objetivo de acelerar

el proceso de entrenamiento y sin afectar significativamente a los resultados (Aggarwal, 2018).

En segundo lugar, es necesario adaptar la distribución de los datos, es decir, los valores de los píxeles, a la entrada de la red. Por ello, las imágenes se normalizaron en el intervalo [-1,1], lo que mejora el proceso de entrenamiento dando lugar a un modelo más estable (Bishop, 1995). Los colores de los píxeles tienen valores comprendidos entre 0 y 255, por lo que es necesario hacer una transformación de cada píxel de la siguiente forma: $valor \times \frac{2}{255} - 1$ obteniendo, de esta forma, un valor entre -1 y 1 (Romero-Oraá *et ál.*, 2019).

3.3. Reducción de la dimensionalidad

La reducción de la dimensionalidad es la transformación de datos de alta dimensión en una representación de dimensionalidad reducida. El objetivo es conseguir que la representación reducida tenga una dimensionalidad que corresponde a la dimensionalidad intrínseca de los datos. La dimensionalidad intrínseca de los datos es el número mínimo de parámetros necesarios que permite distinguir las distintas clases entre sí (Van Der Maaten *et ál.*, 2009). En este contexto, la incrustación de vecinos estocásticos distribuidos en t (*t-distributed Stochastic Neighbor Embedding, t*-SNE) es una técnica de reducción de dimensionalidad no lineal desarrollada por Laurens van der Maaten y Geoffrey Hinton en 2008 a partir del método *Stochastic Neighbor Embedding* de Hinton y Roweis (van der Maaten & Hinton, 2008).

La incrustación de vecinos estocásticos (*Stochastic Neighbor Embedding,* SNE) comienza convirtiendo un espacio euclídeo de distancias entre puntos de datos de alta dimensión, X, a probabilidades condicionadas que representan sus similitudes. En otras palabras, la similitud entre x_i y x_j, en X, es la probabilidad condicionada, $p_{i/j}$, de que x_j escogería a x_i como su vecino. Esto significa que para el espacio Y de la dimensión a la que queremos reducir, se pueden obtener unas probabilidades condicionadas, $q_{i/j}$, similares para los puntos y_i *e* y_j, a partir de los puntos x_i y x_j, de X. El objetivo del algoritmo SNE es encontrar una representación de baja dimensión que minimice la diferencia entre $p_{i/j}$ y $q_{i/j}$ (van der Maaten & Hinton, 2008).

Con el mismo objetivo surgió el algoritmo *t*-SNE. La diferencia entre ambos es que el algoritmo t-SNE es más fácil de optimizar, y, al reducir la posibilidad

de que todas las observaciones se junten en el centro del mapa de visualización, produce una visualización más clara (van der Maaten & Hinton, 2008).

Del mismo modo que SNE, *t*-SNE comienza creando una distribución de probabilidad que represente similitudes entre vecinos, entendiendo similitud del punto de datos x_i con el punto de datos x_j como la probabilidad condicionada $p_{i|j}$ de que x_i elegiría x_j como su vecino (van der Maaten & Hinton, 2008). Posteriormente, para cada punto de datos x_i, se centra una distribución gaussiana sobre ese punto y se mide la densidad de todos los puntos x_j, bajo esa distribución gaussiana. Esto proporciona un conjunto de probabilidades para todos los puntos que son proporcionales a las similitudes entre ellos. La distribución gaussiana o el círculo se pueden manipular a través de un parámetro que se conoce como perplejidad, que influye en la varianza de la distribución y en el número de vecinos más cercanos. Los valores típicos de perplejidad están entre 5 y 50 (van der Maaten & Hinton, 2008). Siguiendo otros estudios, en este TFM se ha empleado un valor de 40 (Chung & Weng, 2017). Denotando la distribución de probabilidad en el espacio original como p_{ij}, la formulación queda (van der Maaten & Hinton, 2008):

$$p_{ij} = \frac{\exp\left(-||x_i - x_j||^2/2\sigma_i^2\right)}{\sum_{k \neq l} \exp\left(-||x_k - x_l||^2/2\sigma_i^2\right)} \tag{3. 1}$$

La segunda parte del algoritmo se centra en el espacio de baja dimensión Y. Se comienza poniendo puntos en el mapa que representan cada punto de la alta dimensión X como un punto en Y. Se fija $y_i \in Y$ correspondiente a $x_i \in X$ y se escoge un área, usando la distribución t de Student con un grado de libertad, con y_i como centro. Esto da un segundo conjunto de probabilidades en el espacio de baja dimensión. La distribución t de Student hace que dos puntos muy alejados en el espacio de alta dimensión estén mucho más alejados en el de baja dimensión, lo que permite un mejor modelado de distancias separadas (van der Maaten & Hinton, 2008). Denotando la distribución de probabilidad en el espacio latente como (van der Maaten & Hinton, 2008):

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq l}(1 + ||y_k - y_l||^2)^{-1}} \tag{3. 2}$$

Por último, se utiliza el conjunto de probabilidades del espacio de baja dimensión (Q_{ij}) para reflejar los datos del espacio de alta dimensión (P_{ij}). Para ello, se mide la diferencia entre las distribuciones de probabilidad de los espacios bidimensionales utilizando la divergencia de Kullback-Liebler (KL) (van der Maaten & Hinton, 2008). La función de coste que minimiza las divergencias de K entre una distribución de probabilidad conjunta, P, en el espacio de alta dimensión y una distribución de probabilidad conjunta, Q, en el espacio de baja dimensión es la siguiente (van der Maaten & Hinton, 2008):

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{3. 3}$$

Usando el método de descenso de gradiente para optimizar la función de costo, la distribución en el espacio original se puede expresar en el espacio de baja dimensión. De esta forma, el gradiente está dado por (van der Maaten & Hinton, 2008):

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + ||y_i - y_j||^2)^{-1} \tag{3. 4}$$

3.4. Redes neuronales

El interés por las redes neuronales surge con el objetivo de emular el comportamiento del cerebro humano (Haykin, 1998). De manera general, las neuronas realizan funciones simples, aunque se suelen modelar como sistemas no lineales y variantes en el tiempo. Comúnmente, los modelos de arquitectura empleados en las redes neuronales agrupan las neuronas en capas conectadas en una estructura paralela. En este contexto, la unidad básica de la arquitectura será la neurona artificial. Una o varias capas compondrán una red neuronal. Por último, una red neuronal o un conjunto de ellas, junto con las interfaces de entrada y salida, constituirán el sistema global de proceso (García Gadañón, 2008).

Las redes neuronales basadas en *deep learning* son un subtipo de esta clase de redes, y dentro de este grupo es donde se encuentra enmarcado el método

desarrollado. Los clasificadores implementados mediante estas redes hacen uso de la entrada original para su entrenamiento, lo que permite prescindir de la etapa previa de extracción de características, así como que el preprocesamiento de los datos de entrada sea más simple que en las redes neuronales convencionales. Asimismo, son redes que cuentan con varias capas ocultas, a diferencia de las redes neuronales convencionales (Schmidhuber, 2014).

3.5. Redes neuronales convolucionales

Las CNNs son un caso especial de redes neuronales de alimentación directa, es decir, los datos atraviesan la red desde la capa de entrada hacia la capa de salida (Aggarwal, 2018). Estas redes son un subtipo de las redes neuronales presentadas anteriormente formadas por neuronas con pesos y sesgos que se ajustan. La arquitectura CNN está diseñada fundamentalmente para que las entradas sean datos bidimensionales, lo que permite codificar ciertas propiedades en la arquitectura (Vieira & Ribeiro, 2018). Esto hace que la función de transferencia sea más eficiente, reduciendo enormemente el número de parámetros y, por lo tanto, que la red sea más fácil de optimizar y menos dependiente del tamaño de los datos. La arquitectura general de las CNN se ilustra en la Figura 3.1, donde se muestran las capas que las constituyen. Esta figura ejemplifica el caso concreto de la clasificación de la severidad de la RD en imágenes de retina.

Otra característica de las CNNs es la utilización de la convolución, en lugar de la multiplicación matricial que emplean las redes neuronales artificiales convencionales. En las redes neuronales convencionales cada neurona está conectada a todas las de la capa siguiente generando un gran número de parámetros de la red. Sin embargo, en las CNNs cada neurona está únicamente conectada a las más cercanas y no a todas las de la capa siguiente (conexiones dispersas), lo que permite reducir significativamente los parámetros de la red a entrenar. Por otra parte, mediante el reparto de los pesos, todas las unidades de una capa usan los mismos pesos y sesgos lo que permite también reducir los parámetros de la red (Aggarwal, 2018; Wang *et ál.*, 2019).

Es posible encontrar más ventajas de las CNN, como la alta disponibilidad de arquitecturas CNN preentrenadas, que pueden emplearse en nuevas tareas de reconocimiento (técnica que se conoce como *transfer learning*), la alta precisión de los resultados que se alcanzan en la clasificación de imágenes y la aplicación de los mismos conocimientos a todas las localizaciones de las señales (Wang *et ál.*,

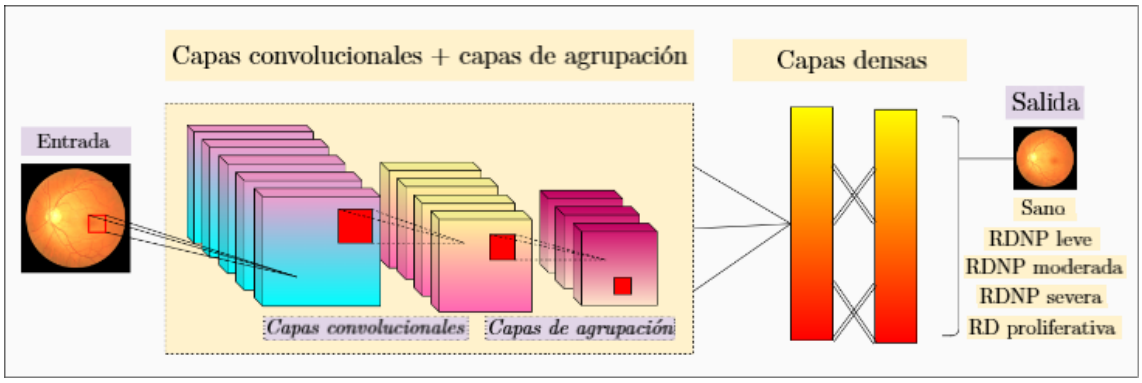


Figura 3.1 Estructura general de una CNN aplicada al problema de clasificación de imágenes de retina.

2019). Esto último significa que, si se traslada la entrada a otra localización dentro de la misma red, las salidas experimentarán una traslación equitativa (Wang *et ál.*, 2019).

En todas las ventajas expuestas anteriormente radica la base de la elección de un modelo basado en CNNs para el sistema planteado en este trabajo.

3.5.1 Capas

Las CNN estándar están compuestas por varias capas que incluyen capas de convolución (*convolution layers*), de agrupamiento (*pooling layers*) y capas completamente conectadas (*fully-conected layers*). Las capas *fully connected* dan cómo salida la probabilidad de pertenecer a cada clase. La salida con mayor probabilidad se decide como la clase seleccionada por el modelo para esa entrada. Todas estas se han empleado en el desarrollo del método presentado.

3.5.1.1 Capa convolucional

Se trata de la capa esencial de una CNN. La capa de convolución usa un núcleo convolucional como filtro de entrada, con el objetivo de convertir la entrada en una representación de nivel más abstracto. El funcionamiento se basa en realizar una convolución entre cada región de la matriz de entrada y el *kernel* o filtro, almacenando los resultados en lo que se denomina mapas de activación (Aggarwal, 2018). Hay que tener en cuenta que, en realidad, no se aplica un solo *kernel*, sino que habrá varios (Aggarwal, 2018). Se distinguen los siguientes parámetros en el proceso de convolución (Emmert-Streib *et ál.*, 2020):

- **Tamaño del *kernel* o filtro (N).** Parámetro que define el tamaño de ventana del filtro, determinando así el tamaño de la región de la imagen de entrada con la que se realiza la convolución. En la Figura 3.2 se puede observar como el área roja de la entrada tiene el mismo tamaño que el filtro mostrado en azul.

- ***Stride* o paso (S).** Parámetro del filtro que modifica el número de píxeles que el *kernel* se desplazará sobre la matriz de píxeles de entrada. Si el *stride* se establece en 1, el filtro se desplaza un píxel por cada convolución. Sin embargo, no es necesario realizar la convolución en cada posición espacial en la capa. Por ejemplo, cuando el *stride* es dos, se desplazan los filtros dos píxeles en cada convolución y así sucesivamente. Como resultado, el uso de *strides* producirá una reducción de la dimensión espacial de la imagen. Esto resulta especialmente útil debido a que no se aplica un único filtro. Si, por ejemplo, realizásemos la convolución con 16 filtros, se obtendrían 16 matrices de salida, cada una de ellas del mismo tamaño que la imagen de entrada. Si suponemos que la imagen tiene una resolución de 64x64 píxeles, tendríamos un total de 65536 neuronas en la salida de dicha capa. El aumento del número de neuronas afecta significativamente a la velocidad de procesamiento de la CNN, por lo que, en ocasiones resulta necesario emplear un *stride* distinto a 1.

- ***Padding* (P).** Parámetro que define el número que píxeles que se añaden en el borde de una imagen. Hay imágenes donde hay información importante lejos del centro de la imagen, específicamente en los bordes. Al realizar la convolución, el *kernel* pasa muy pocas veces por las esquinas de la imagen comparado con las veces que pasa por el centro de la imagen, lo que se traduce en una pérdida de información. Por ello, se añaden más pixeles alrededor de la imagen, de manera que toda información se sitúe más cerca del centro. Otro de los usos de este parámetro es evitar que la matriz de la imagen resultante sea muy pequeña. En la Figura 3.2 se puede observar que con una imagen de entrada 6 x 6 y un *kernel* 3 x 3, se obtiene como resultado una matriz 4 x 4. Como en una CNN hay varias capas convolucionales, cada vez que la matriz pase por estas capas perderá

tamaño. En este contexto, existen dos opciones de *padding* (Aggarwal, 2018):

- Rellenar la imagen con ceros, lo que se conoce como *zero-padding*.
- Eliminar la parte de la imagen donde el filtro no encaja. Esto se llama relleno válido que mantiene solo una parte válida de la imagen.

Estos tres parámetros son los más utilizados para controlar el volumen de salida de una capa convolucional. Específicamente, para una entrada de dimensión $W_{entrada} \times H_{entrada} \times Z$, donde Z es el número de canales de la imagen (3 si es en color), la dimensión del mapa de activación, es decir, $W_{salida} \times H_{salida} \times D$ puede ser calculado a través de (Emmert-Streib *et ál.*, 2020):

$$W_{salida} = \frac{(W_{entrada} - N + 2P)}{S + 1}$$
$$H_{salida} = \frac{(H_{entrada} - N + 2P)}{S + 1} \quad (3.\ 5)$$
$$D = Z$$

En la Figura 3.2 se ilustra un ejemplo del funcionamiento de la capa convolucional. Considerando un tamaño de *kernel* de 3x3, un *stride* de 1 y un *padding* de 0, entonces el filtro de tamaño 3 x 3, se desplazará un píxel en cada iteración de izquierda a derecha comenzando por la posición superior izquierda. Cuando el filtro llegue al límite derecho, empezará de nuevo el proceso de convolución desde la segunda fila y así sucesivamente hasta cubrir por completo

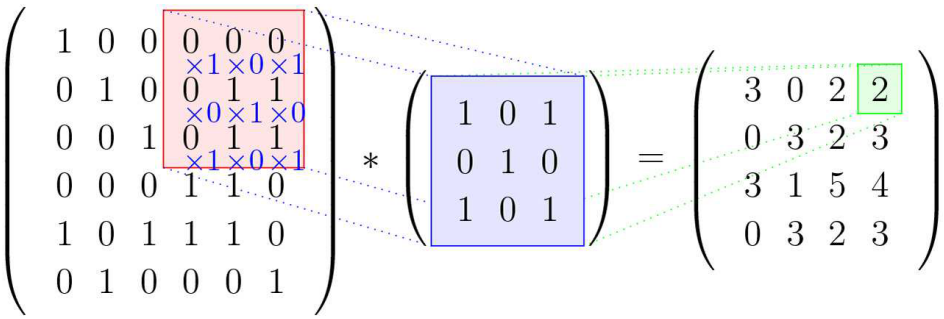


Figura 3.2 Ejemplo de cálculo del mapa de activación (Fuente: Emmert-Streib *et ál.*, 2020).

la matriz de píxeles de entrada. El área en rojo indica el parche local que se va a convolucionar con el *kernel*, y el resultado se almacena en el campo verde del mapa de activación (Emmert-Streib *et ál.*, 2020).

En este trabajo, los valores de los parámetros mencionados anteriormente dependen únicamente de la arquitectura empleada, ya que no se han añadido más capas convolucionales al modelo.

3.5.1.2 Capa de agrupación o pooling

La capa de agrupación o *pooling* busca que las salidas conseguidas tras cada capa de convolución no se vean afectadas por pequeñas variaciones en las entradas. Para conseguir esto se sustituye el valor de cada unidad de salida de la capa convolucional por otro valor que tenga también en cuenta el valor de las salidas más próximas. Existen diferentes tipos de *pooling*, siendo los más comunes (Aggarwal, 2018; Fronzetti, 2019):

- ***max-pooling***. Cada unidad toma el valor máximo entre todas las que la rodean.
- ***average-pooling***. Cada unidad toma el valor medio entre todas las que la rodean.

En la Figura 3.3 se muestra un ejemplo de esta operación, en la que se combinan las etapas de *pooling* (*average-pooling* en la parte izquierda y *max-pooling* en la parte derecha).

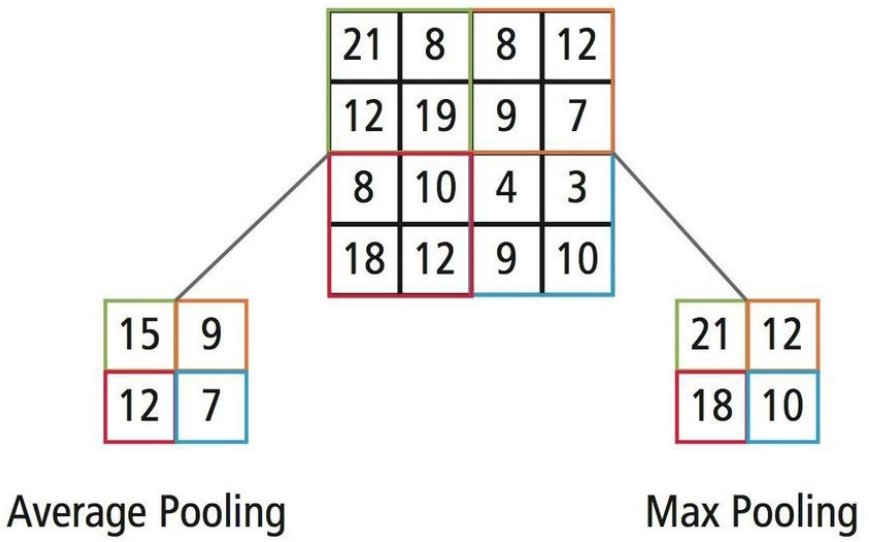


Figura 3.3 Ejemplos de operación de las capas *max pooling* y *average pooling* (Fuente: Fronzetti, 2019).

3.5.1.3 Capa densa o fully connected

Las capas densas o *fully connected* son capas de cálculo que conectan cada neurona en una capa con todas las salidas de la capa anterior (Aggarwal, 2018). Se usan como capas finales en problemas de clasificación, donde la última de ellas contiene tantos nodos como clases entre las que clasificar existan. Al igual que las capas convolucionales, hacen uso de funciones de activación no lineales (Aggarwal, 2018).

3.5.2 Funciones de activación

De forma general, a los valores de las operaciones de convolución entre el filtro y la entrada, se aplica una función de activación no lineal. El uso de funciones de activación no lineales permite que la red sea capaz de modelar funciones de gran complejidad, haciendo que la salida no sea una combinación lineal simple de la entrada (Emmert-Streib *et ál.*, 2020). De esta manera, se consigue una semejanza con el funcionamiento biológico del cerebro, dónde las neuronas se activan al recibir un estímulo que reúne ciertas características (Xu *et ál.*, 2015).

En este contexto, existen diferentes tipos de funciones de activación, siendo algunas de las más habituales las que se detallan a continuación (Feng *et ál.*, 2019; Zhou *et ál.*, 2019):

- **Sigmoidea.** Esta no linealidad se define como:

$$\sigma(x) = \frac{1}{1+e^{-x}}, \quad x \epsilon R \tag{3. 6}$$

  Se puede observar que σ (x) $\in$ (0, 1) para todo x $\in$ R. Además, σ es una función monótona creciente, $\lim_{x\to\infty} \sigma(x) = 1$ y $\lim_{x\to-\infty} \sigma(x) = 0$.

  Esto hace que la función de activación sigmoidea sea adecuada cuando el objetivo es producir salidas contenidas en el rango [0, 1], como probabilidades o imágenes normalizadas. También se puede demostrar que $\lim_{x\to\infty} \sigma'(x) = \lim_{x\to-\infty} \sigma'(x) = 0$.

  Sin embargo, cuenta con la desventaja de que, si el número de capas de la red es elevado, el valor del gradiente de dicha función disminuirá gradualmente y el aprendizaje de la red será muy lento.

- **Tanh.** La función tangente hiperbólica de sigmoide (*tanh*) se define como:

$$tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}, x \in R \tag{3. 7}$$

Se puede demostrar que tanh(x) ∈ (−1, 1) para todo x ∈ R. Además, tanh es una función monótona creciente, $\lim_{x\to\infty} \tanh(x) = 1$ y $\lim_{x\to-\infty} \tanh(x) = -1$. Similar a la no linealidad sigmoidea, esta no linealidad puede conducir a gradientes de fuga y rara vez se usa en capas intermedias de CNN.

- **ReLU.** La función conocida como unidad lineal rectificada (*Rectified Linear Unit*, ReLU) se define como:

$$ReLU(x) = max(0, x), x \in R \tag{3. 8}$$

Se puede ver que ReLU'(x) = 1 para x> 0 y que ReLU'(x) = 0 para x <0. La no linealidad de ReLU generalmente conduce a una convergencia más rápida en comparación con las no linealidades sigmoidea o *tanh*, y generalmente funciona bien en CNN con una estrategia de inicialización de peso y una tasa de aprendizaje elegidas adecuadamente.

- **Leaky ReLU.** La función ReLU con fugas se define utilizando un parámetro adicional α ∈ (0, 1):

$$\emptyset(v) = \begin{cases} \alpha \cdot v & v \leq 0 \\ v & en\ otro\ caso \end{cases} \tag{3. 9}$$

Aunque α es un hiperparámetro elegido por el usuario, también es posible aprenderlo.

Esta variante soluciona el problema de la función de activación ReLU, que consiste en la anulación de un gran número de neuronas de la red como consecuencia de que todos los valores negativos proporcionen una salida nula.

En la Figura 3.4 se muestran todas las funciones descritas anteriormente.

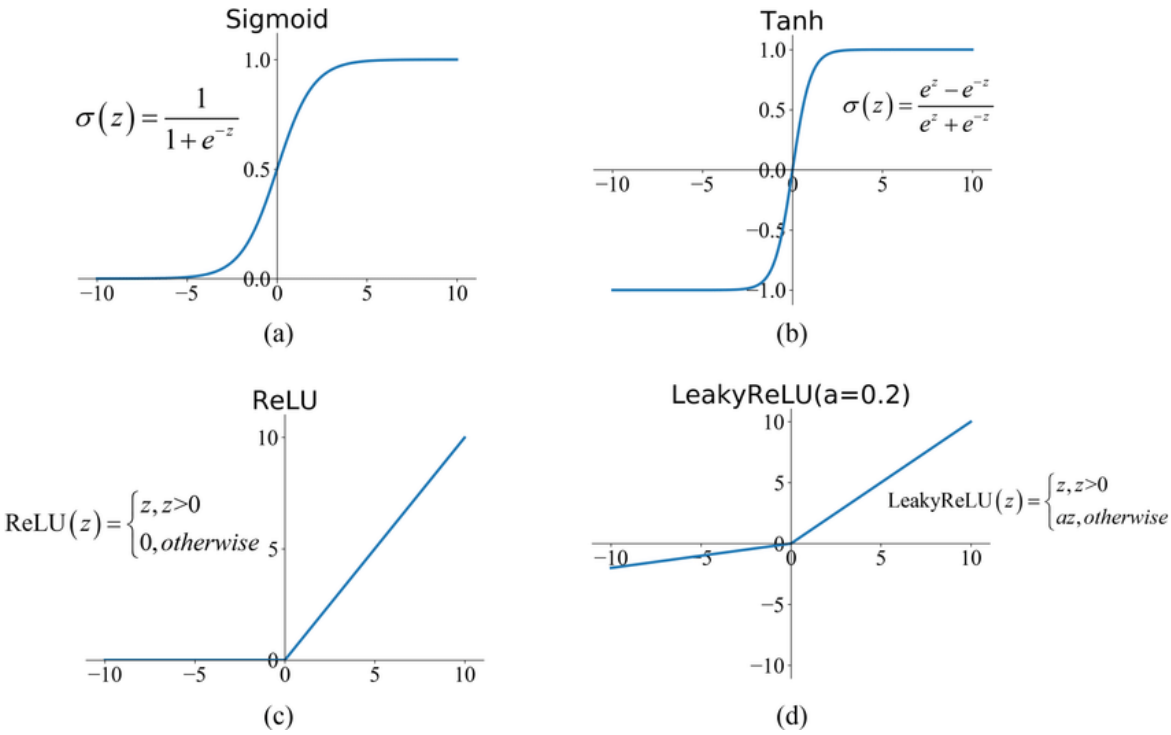


Figura 3.4 Funciones de activación comúnmente usadas. **(a)** Sigmoidea**. (b)** Tanh. **(c)** ReLU. **(d)** Leaky ReLU. (Fuente: Feng *et ál.*, 2019).

3.6. Arquitecturas CNN

El concepto de CNN hace referencia a redes neuronales con una estructura que es invariable a las operaciones de rotación y traslación. La arquitectura CNN está adaptada para la manipulación de datos bidimensionales, lo que permite codificar ciertas propiedades en la arquitectura. Esto hace que la función de transferencia sea más eficiente, reduciendo enormemente el número de parámetros y, por lo tanto, hace que la red sea más fácil de optimizar en problemas de procesado de imagen (Vieira & Ribeiro, 2018).

Hoy en día, las CNN se consideran los algoritmos más utilizados entre las técnicas de inteligencia artificial. La historia de las CNN comienza con los experimentos neurobiológicos realizados por Hubel y Wiesel, en los años 1959-1962. Su trabajo proporcionó una plataforma para muchos modelos cognitivos, y la CNN sustituyó a casi todos ellos. A lo largo de las décadas, se han realizado diferentes esfuerzos para mejorar el rendimiento de las CNN. La historia evolutiva de las arquitecturas de las CNN profundas se representa en la Figura 3.5 (Khan *et ál.*, 2020).

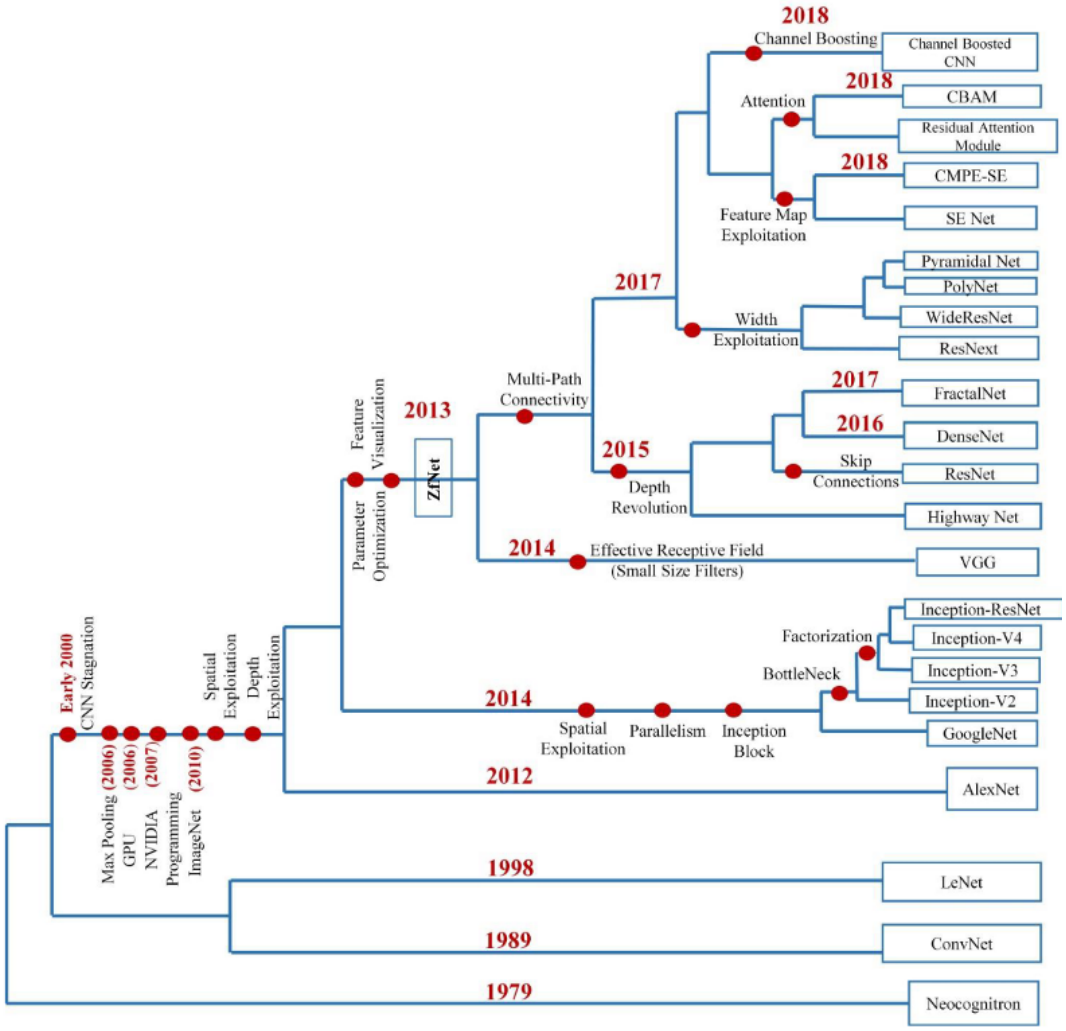


Figura 3.5 Historia evolutiva de las CNN profundas que muestra las innovaciones arquitectónicas hasta las arquitecturas actuales (Fuente: Khan *et ál.*, 2020).

La disponibilidad de amplios datos de entrenamiento y los avances en el hardware son los factores que han contribuido al avance en la investigación de las CNN. Pero las principales fuerzas motrices que han acelerado la investigación y han dado lugar al uso de las CNN en tareas de clasificación y reconocimiento de imágenes son las estrategias de optimización de parámetros y las nuevas ideas arquitectónicas (Gu *et ál.*, 2018; Sinha *et ál.*, 2017). El principal avance en el rendimiento de las CNN lo aportó AlexNet, que mostró un rendimiento ejemplar en 2012, reduciendo la tasa de error de un 25.8% a 16.4%, en comparación con las redes anteriores (Krizhevsky *et ál.*, 2012).

En general, se observa que las mejoras significativas en el rendimiento de las CNN se produjeron entre 2015 y 2019. La capacidad de aprendizaje de una CNN

suele depender de su profundidad. Además, el algoritmo más utilizado para entrenar redes neuronales es el descenso del gradiente: el gradiente es un cálculo que nos permite saber cómo ajustar los parámetros de la red de tal forma que se minimice el error a la salida. El problema es que, a medida que aumenta la profundidad de la red, el gradiente se irá desvaneciendo a valores muy pequeños, impidiendo el aprendizaje de la red. En el caso peor, esto puede impedir que la red neuronal continúe su entrenamiento. En este contexto, diferentes estudios demostraron que las arquitecturas como VGG, ResNet, ResNext, DenseNet, etc., disminuían el problema de descenso de gradiente, al mismo tiempo que proporcionaban buenos resultados en problemas de reconocimiento y clasificación (Khan *et ál.*, 2020).

Por otra parte, cabe mencionar que la arquitectura de GoogLeNet fue la primera en abordar el problema de los altos recursos que consume entrenar una red convolucional. A medida que las redes neuronales se hacen más profundas, el consumo de memoria aumenta significativamente. Para solventar este problema, GoogLeNet fue una de las primeras arquitecturas que introdujo la idea de que las capas de CNN no siempre tienen que apilarse de manera secuencial. Los autores del documento mostraron que se puede mejorar el rendimiento aumentando también la anchura de la red. A partir de esta arquitectura, se desarrollaron los modelos Inception-V2, Inception-V3, Inception-V4 e Inception-ResNet (Szegedy *et ál.*, 2014, 2015, 2017).

En este trabajo, con el objetivo de encontrar la arquitectura más adecuada para clasificar el grado de severidad de la RD mediante retinografías, se estableció una comparación entre 4 arquitecturas CNN distintas: ResNet-50, Inception-V3 DenseNet-201 y MobileNet-V2. Las arquitecturas seleccionadas han tenido éxito en problemas de clasificación de imagen y muchas de ellas se han utilizado con imágenes de fondo de ojo *(*Dekhil *et ál.*, 2019*;* Kassani *et ál.*, 2019*;* Sengupta *et ál.*, 2020). En esta sección se presenta una breve descripción de cada una de ellas.

3.6.1 ResNet

La arquitectura ResNet fue propuesta en 2015 por investigadores de *Microsoft Research* como un intento de crear un modelo de clasificación de imágenes que solucionara los problemas de las redes convencionales, cuya respuesta, al contrario de lo que se podía intuir, se degradaba al aumentar la profundidad de la red. El propósito era solucionar el problema de desvanecimiento del gradiente. El

gradiente representa la velocidad a la que se actualizan los pesos, por lo que gradientes muy pequeños implican que los pesos de las capas lejanas de la capa de salida apenas se actualizan. Esto significa que casi no se produce aprendizaje para las capas cercanas a la capa de entrada (Aggarwal, 2018). La disminución del valor del gradiente es lo que se conoce como desvanecimiento del gradiente. Por ello, aumentar el número de capas no mejoraba el rendimiento de la red (He *et ál.*, 2015).

Este tipo de arquitectura conocida como *Residual Network* se caracteriza por emplear conexiones de salto, que permiten evitar el entrenamiento de algunas capas, conectándose directamente a la salida. Esto presenta la ventaja de que, si alguna capa afecta negativamente al rendimiento de la red, es posible evitar la influencia de tal capa (He *et ál.*, 2015).

ResNet-50 ha sido la primera arquitectura empleada en este trabajo. Se trata de una red neuronal que cuenta con 50 capas, siendo 48 capas de convolución, una capa *max pool* y una capa *average pool* como se muestra en la Figura 3.6 (He *et ál.*, 2015). La decisión de utilizar esta arquitectura radica en que ha sido empleada con éxito en tareas de clasificación de la RD (Adriman *et ál.*, 2021; Kassani *et ál.*, 2019).

3.6.2 Inception-V3

La segunda arquitectura empleada ha sido Inception-V3. Inception-V3 es la tercera generación de arquitecturas de CNNs de Inception, que destaca por el uso de bloques de normalización por lotes (Szegedy *et ál.*, 2014). De manera general, la forma más simple de mejorar el rendimiento de la red es aumentar la profundidad y el ancho de la misma, lo que significa aumentar el número de capas

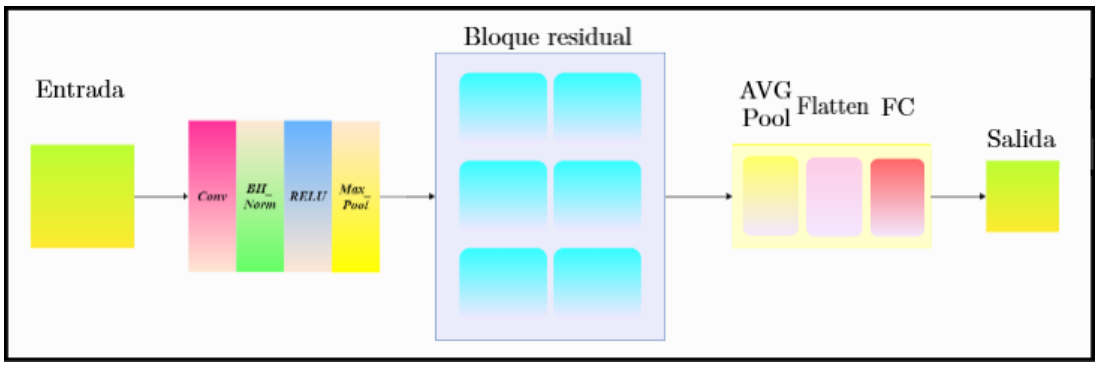


Figura 3.6 Arquitectura ResNet-50 (adaptado de Hattiya *et ál.* (2021)).

ocultas, así como el número de neuronas en cada capa. Sin embargo, esta técnica presenta algunos problemas (Szegedy *et ál.*, 2014, 2015):

- El espacio de parámetros que se obtiene es grande y fácil de sobreajustar.
- Los recursos informáticos necesarios para el entrenamiento de la red son mayores.
- A medida que aumenta la profundidad de la red, más difícil se hace la optimización de la misma.

Con estas premisas, el objetivo de los autores de la red Inception fue mejorar la utilización de los recursos informáticos de la red, y aumentar el ancho y la profundidad de la red, mientras que la cantidad de cálculo permanecía sin cambios (Szegedy *et ál.*, 2014). Inception-V3 emplea la factorización de convoluciones para reemplazar las convoluciones de dimensiones superiores a 3 x 3 por combinaciones de convoluciones de dimensión inferior. Esta factorización permite reducir el coste computacional. Por ejemplo, los autores explican que las convoluciones de 5 x 5 tienen un coste computacional 2.79 veces superior a las de 3 x 3 (Szegedy *et ál.*, 2015).

La arquitectura Inception-V3 posee 48 capas de profundidad, con convoluciones factorizadas de 7x7 (combinaciones de convoluciones de dimensión inferior a 3 x 3), optimizador RMSProp para reducir el error cometido por la red y la normalización por lotes, como se observa en la Figura 3.7. También utiliza el suavizado de etiquetas para evitar el sobreajuste de la red (Szegedy et al., 2015). El suavizado de etiquetas es una técnica de regularización que introduce ruido en las etiquetas. Muchas anotaciones manuales son el resultado de varios participantes, que pueden seguir diferentes estándares e incluso cometer algunos errores. Esta técnica tiene en cuenta el hecho de que los conjuntos de datos pueden

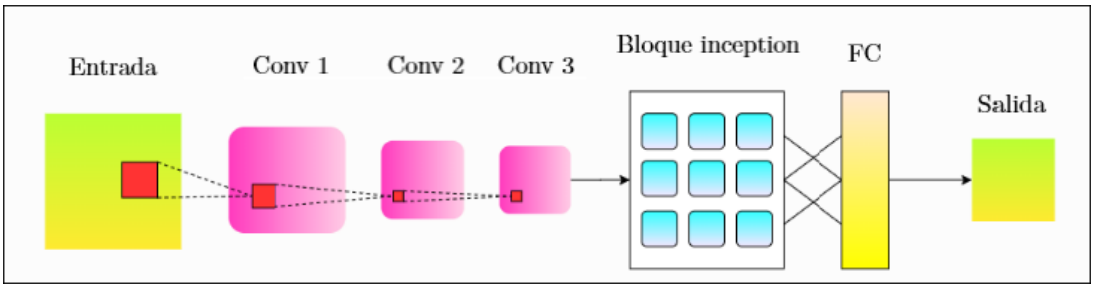


Figura 3.7 Arquitectura Inception-V3 (adaptado de Hattiya *et ál.* (2021)).

contener errores, por lo que maximizar la probabilidad de forma directa puede ser perjudicial (Müller *et ál.*, 2019).

Esta arquitectura ha sido empleada por diversos autores para la identificación de las diferentes etapas de la RD (Bhardwaj *et ál.*, 2021; Masood *et ál.*, 2017), hecho que ha motivado la utilización de la misma.

3.6.3 DenseNet

La arquitectura DenseNet se propuso en el año 2016 para solventar el problema de desvanecimiento del gradiente, en el mismo contexto que la arquitectura ResNet. El problema que arrastraba esta última era que preservaba explícitamente la información de manera que, muchas capas podían contribuir con ninguna o muy poca información.

Algunos estudios demostraron que las CNNs podían ser sustancialmente más profundas, precisas y eficientes de entrenar si contienen conexiones más cortas entre las capas cercanas a la entrada y las cercanas a la salida. En este contexto, surgió la CNN densa DenseNet, que conecta cada capa a cada una de las otras capas de una manera *feed-forward*. Mientras que las CNNs tradicionales con L capas tienen L conexiones: una entre una capa y su siguiente capa, las redes DenseNet tienen $\frac{L(L+1)}{2}$ conexiones directas. En concreto, la red DenseNet-201, empleada en este trabajo, cuenta con 201 capas de profundidad (Huang *et ál.*, 2016). Su arquitectura se puede observar en la Figura 3.8.

Distintos autores han utilizado esta arquitectura en la tarea de clasificación de la severidad de la RD (L. Li et al., 2020). Por este motivo, se ha decidido emplear dicha arquitectura CNN.

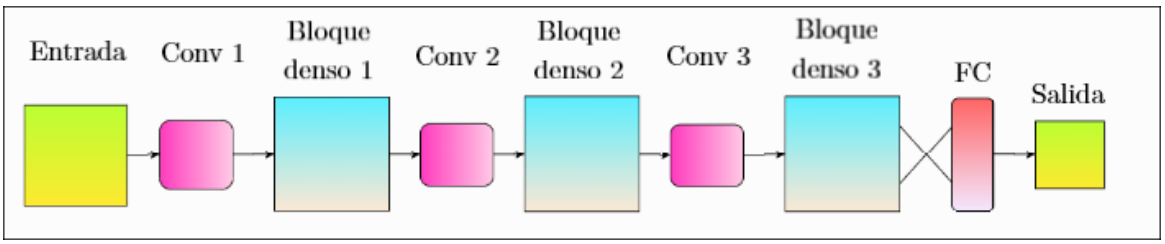


Figura 3.8 Arquitectura DenseNet-201 (adaptado de Hattiya *et ál.* (2021)).

3.6.4 MobileNet

La arquitectura MobileNet fue propuesta en 2017 por investigadores de *Google*. Se basa en una arquitectura racionalizada que utiliza convoluciones separables en profundidad para construir redes neuronales profundas de poco peso. Este tipo de convoluciones permiten aplicar un único filtro en cada canal de entrada, a diferencia de una convolución estándar en la que los filtros se aplican a todos los canales de entrada. Se trata de una arquitectura formada por 28 capas, que aplica la normalización de *batch* tras cada capa de convolución con la función de activación ReLU (Howard *et ál.*, 2017). Esta arquitectura introduce dos hiperparámetros globales que compensan eficazmente la latencia y la precisión. Estos hiperparámetros permiten elegir el tamaño adecuado para la aplicación en función de las restricciones del problema. Las diferentes capas que componen la arquitectura MobileNet se muestran en la Figura 3.9.

A principios de 2018 surgió una nueva versión de este tipo de arquitectura, llamada MobileNet-V2. En cuanto a su estructura, cuenta con una capa inicial convolucional formada por 32 filtros seguida de 19 capas de cuello de botella residuales (Sandler *et ál.*, 2018). La arquitectura MobileNet-V2 también ha sido empleada con éxito en la clasificación de las distintas etapas de la RD en varios estudios (Patel & Chaware, 2020; Sheikh & Qidwai, 2020), motivando la elección de la misma en este trabajo.

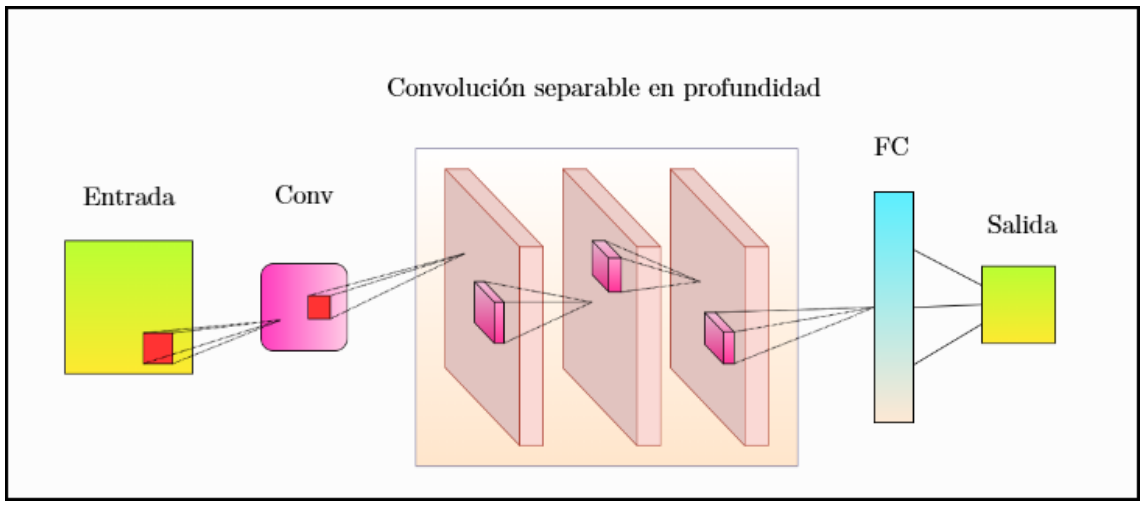


Figura 3.9 Arquitectura MobileNet (adaptado de Hattiya *et ál.* (2021)).

3.7. Modelo desarrollado

Una vez explicados los conceptos básicos del funcionamiento y de la estructura de las redes neuronales profundas, se procede a explicar los detalles de la arquitectura y de las técnicas que se han empleado en este trabajo. El sistema desarrollado se trata de un clasificador automático que permite establecer el grado de severidad de la RD en imágenes de fondo de ojo.

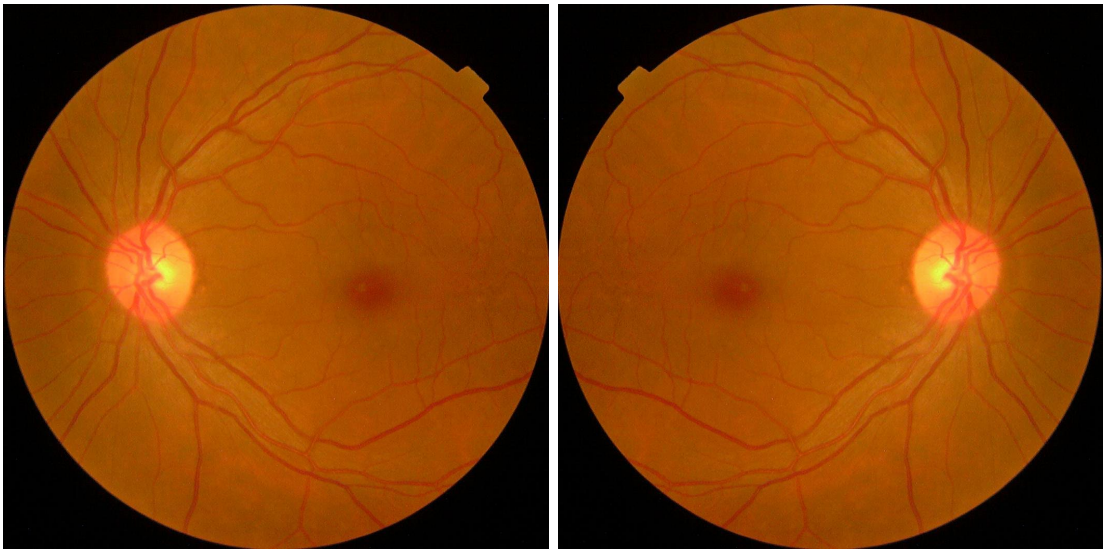
3.7.1 Data augmentation

Las redes neuronales profundas se benefician de grandes cantidades de datos de entrenamiento, y es una práctica común ampliar artificialmente el conjunto de datos mediante un proceso llamado aumento de datos o *data augmentation* (Aggarwal, 2018). Esta técnica permite crear imágenes sintéticas a partir de las imágenes disponibles. No obstante, se si eligen transformaciones que generan imágenes irreales, el entrenamiento se vería afectado negativamente, provocando que la red no aprendiese las características de las imágenes reales (Aggarwal, 2018).

En general, la elección de la transformación utilizada para el aumento de datos depende del conjunto de los mismos, pero hay algunas estrategias comunes para el aumento de datos en el campo de visión artificial (Zhou *et ál.*, 2019):

- **Voltear**. La imagen x se refleja en una o dos dimensiones, produciendo una o dos muestras adicionales. Se puede ver un ejemplo de esta transformación en la Figura 3.10.
- **Rotación aleatoria**. Una imagen x se gira mediante un ángulo aleatorio ϕ.

Además, en este trabajo, se utilizó el aumento de datos en tiempo real, lo que significa que en cada iteración se generó un *batch* de imágenes nuevo. Por tanto, el número de imágenes que se empleadas en el entrenamiento dependió del número de épocas. Las transformaciones que se aplicaron sobre las 2930 imágenes originales del conjunto de entrenamiento fueron volteos horizontales y verticales aleatorios, siguiendo el trabajo de Pratt *et ál.* (2016). Con estas operaciones se consigue aumentar el conjunto de imágenes de entrenamiento, aumentando la precisión del método y evitando el sobreentrenamiento. Es decir, se evita que el modelo se ajuste a unas características muy específicas de los datos de



(a) **(b)**

Figura 3.10 (a) Imagen original. **(b)** Imagen volteada horizontalmente.

entrenamiento y no responda adecuadamente sobre los conjuntos de datos de test (Aggarwal, 2018).

3.7.2 Transfer-learning y fine-tuning

La técnica *transfer learning* consiste en que las redes se entrenan previamente en grandes conjuntos de datos disponibles y posteriormente, la red se ajusta a una cantidad limitada de datos dependientes de la aplicación, de modo que las redes se adaptan a la aplicación actual. Es muy popular en la creación de sistemas basados en *deep learning* porque permite entrenar redes neuronales profundas con relativamente pocos datos. De esta forma, no resulta necesario disponer de un conjunto de imágenes tan amplio como en el caso de tener que entrenar la red desde cero. La forma más habitual de aplicar esta técnica es el uso de redes pre-entrenadas, por ello, esta ha sido la forma empleada en el desarrollo de este trabajo (Shao *et ál.*, 2015).

Para cada una de las arquitecturas consideradas en este trabajo, se partió de los pesos obtenidos con imágenes del proyecto ImageNet. Esta base de datos se compone de más de 14 millones de imágenes pertenecientes a más de 20000 clases. Además, cada clase está compuesta por varias subclases. Por ejemplo, la clase "animal" es la más extensa con 3822 subclases, entre las que se pueden encontrar "perro", "gato" o "león" entre otras. Otras clases son "comida",

"electrodomésticos" o "plantas" (Russakovsky et al., 2014). Esta base de datos se ha utilizado previamente con éxito en la clasificación de imágenes de fondo de ojo (Choi *et ál.*, 2017).

Además de la utilización de redes pre-entrenadas, se aplicó la técnica *fine-tuning*, asociada al concepto de *transfer learning*. En los modelos basados en redes profundas, las primeras capas permiten identificar las características generales de más bajo nivel aplicables a todas las imágenes, como bordes o patrones; mientras que las últimas capas identificarían las características más relevantes para el problema que se quiere resolver, como la presencia de exudados o hemorragias en retinografías. La técnica *fine-tuning* permite reentrenar todas o algunas de las capas finales de la CNN utilizando imágenes específicas del problema a resolver. De esta manera, el método resultante es más preciso en comparación con los modelos entrenados desde cero, ya que los pesos se adaptan mejor al objetivo del problema (Too *et ál.*, 2019). En este trabajo, se reentrenaron las últimas cinco capas para todas las arquitecturas empleadas (He *et ál.*, 2015).

3.7.3 Dropout

La regularización, en general, abarca un amplio número de técnicas destinadas a reducir el sobreajuste de los modelos basados en redes neuronales. Para evitar el sobreentrenamiento de los modelos, en este trabajo se ha empleado la técnica *dropout*. El objetivo de aplicar esta técnica es cancelar algunas de las neuronas a medida que se entrena la red, de manera que la las neuronas canceladas vuelvan a aprender sin sobreajuste. La eliminación de una neurona implica también la eliminación de las conexiones de entrada y salida asociadas a esa neurona (Srivastava *et ál.*, 2014). La elección de las neuronas a anular es aleatoria. Durante el entrenamiento, *dropout* se implementa manteniendo una neurona activa con cierta probabilidad $p \in (0,1)$ siendo fijo este valor de p durante toda la fase de entrenamiento. Por ejemplo, si se asigna un *dropout* entre la capa 5 y 6 del modelo, y se establece un valor de 0.3 para el parámetro *p,* estamos diciendo que todos los valores que van a salir de la capa 5 en tiempo de entrenamiento tienen un 30% de probabilidad de llegar a la capa 6, mientras que el resto serán forzados a ser 0. Al eliminar de manera aleatoria valores entre las capas, se pueden crear modelos grandes sin que esto impacte la capacidad del modelo de generalizar a valores no vistos en el entrenamiento (Achille & Soatto, 2018; Srivastava *et ál.*, 2014).

Concretamente, en este trabajo se ha empleado un índice de *dropout* de 0.5 en la fase de entrenamiento, siguiendo otros estudios (Kandel & Castelli, 2020; Shaban *et ál.*, 2020).

3.7.4 Arquitectura empleada

Con el objetivo de encontrar la arquitectura más adecuada para clasificar el grado de severidad de la RD en retinografías, se estableció una comparación entre cuatro arquitecturas CNN distintas: ResNet-50, Inception-V3, DenseNet-201 y MobileNet-V2.

Para adaptar estas arquitecturas al problema de clasificación multiclase, se han añadido una capa *average pooling*, y dos capas densas de 2018 y 5 neuronas, respectivamente (Kandel & Castelli, 2020). Además, se empleó una combinación del método *dropout*, junto con la capa *average pooling*, como técnica de regularización. Se ha demostrado que esto permite reducir el sobreajuste de las CNN (Aggarwal, 2018). El número de neuronas de la última capa se corresponde con el número de clases que queremos discriminar, siendo estas: paciente sano, paciente con RDNP leve, paciente con RDNP moderada, paciente con RDNP severa y paciente con RD proliferativa. En la primera capa densa se utilizó una función de activación tipo ReLU, ya que conduce a una convergencia más rápida en comparación con las no linealidades sigmoidea o *tanh*. Además, esta función de activación ha sido empleada por otros autores para la clasificación de la severidad de la RD (Mohammadian *et ál.*, 2017). Por otra parte, en la última capa densa se utilizó la función de activación *softmax*. Esta función de activación se emplea en tareas de clasificación multiclase, devolviendo una distribución de probabilidad entre las salidas, de forma que cada una de estas salidas indica la probabilidad de que la imagen de entrada pertenezca a dicha clase (Aggarwal, 2018).

Además, en la fase de *fine-tuning*, se empleó la técnica *early-stopping*. De esta manera, se puede detectar la convergencia del entrenamiento, siendo útil para evitar el sobreentrenamiento. Esta técnica permite el ahorro de tiempo de ejecución ya que una vez detectado puede detenerse el entrenamiento y no es necesario llegar al máximo de épocas. Se aplicó la entropía cruzada categórica como función de pérdida al tratarse de un problema de clasificación multiclase y Adam (*Adaptive moment estimation*) como algoritmo de optimización siguiendo otros estudios (Alyoubi *et ál.*, 2021; Mohammadian *et ál.*, 2017).

A continuación, se exponen los hiperparámetros seleccionados para el modelo desarrollado, que fueron comunes para cada una de las arquitecturas consideradas. Los hiperparámetros son las variables que determinan la estructura de la red y cómo se entrena la red. Estos hiperparámetros se definen antes de entrenar el modelo (Aggarwal, 2018). En este TFM se han ajustado los hiperparámetros tasa de aprendizaje y tamaño de lote o *batch*. En general, tiene sentido elegir el tamaño del lote o *batch* lo más grande posible dada la arquitectura de red y el tamaño de la imagen, y posteriormente, elegir la mayor tasa de aprendizaje posible que permita un aprendizaje estable (Le *et ál.*, 2017). Si el error oscila, en lugar de disminuir, se recomienda reducir la tasa de aprendizaje inicial (Le *et ál.*, 2017). También es común cambiar la tasa de aprendizaje durante el entrenamiento dependiendo del número actual de épocas. Al dividir la BD en *batches,* cuando se termina de procesar cada partición o *batch* en la fase de entrenamiento se dice que ha concluido una época. A mayor tamaño de *batch*, más rápido se entrenará el modelo. Pero si este tamaño es demasiado grande, necesitará más épocas de entrenamiento (Le *et ál.*, 2017). Siguiendo otros trabajos, la tasa de aprendizaje se fijó en 0.0001 (Alyoubi et al., 2021; Kassani et al., 2019). Del mismo modo, para evitar sobreentrenamiento en épocas avanzadas, se redujo la tasa de aprendizaje en un factor de 0.5 cada vez que el error de validación alcanzase un mínimo y se mantuviese constante (Majumder & Kehtarnavaz, 2021).

Finalmente, para iniciar el proceso de entrenamiento, era esencial encontrar el tamaño de *batch* que se ajustara a la RAM y a la GPU, respectivamente. Este hiperparámetro indica el número de imágenes que pueden leerse en la RAM para el proceso de entrenamiento. Todo el proceso de entrenamiento se realizó utilizando una GPU NVIDIA GeForce GTX 960M de 2GB de RAM alojada en un ordenador con procesador Intel core i7 y 16GB de RAM, lo que ha condicionado el uso de un tamaño de *batch* de 8 imágenes.

3.8. SHapley Additive exPlanations

La inteligencia artificial y, más concretamente, el *machine learning* y el *deep learning* aplicado a la ayuda al diagnóstico de diferentes patologías ha sido un área de investigación de alta productividad científica durante las últimas décadas (Shin *et ál.*, 2016; Tjoa & Guan, 2020). En los últimos años, ha surgido en la sociedad una demanda para que las decisiones automáticas tomadas por estos algoritmos sean explicables y, por tanto, justificables ante los usuarios finales de

los mismos, como es el caso de los pacientes de los sistemas de salud (Došilović *et ál.*, 2018; Tjoa & Guan, 2020). En este sentido, bajo el término *explainable artificial intelligence* (XAI) ha surgido también una nueva familia de técnicas que dan respuesta a esta demanda, a la vez que pueden ser usadas para controlar y mejorar los modelos de *deep learning* e, incluso, revelar nuevo conocimiento sobre la temática o problema bajo estudio (Došilović *et ál.*, 2018).

Englobado dentro de las técnicas XAI, se encuentra *SHapley Additive exPlanations* (SHAP). SHAP es un enfoque de la teoría de juegos que permite explicar el resultado de cualquier modelo de aprendizaje automático. Proporciona visualizaciones intuitivas e interactivas que muestran qué características son más relevantes para una determinada predicción y para el modelo en su conjunto (Lundberg & Lee, 2017).

SHAP asigna a cada característica un valor de importancia para una predicción particular, incluyendo resultados teóricos que muestran que existe una solución única con un conjunto de propiedades deseables (Lundberg & Lee, 2017). Matemáticamente, la definición de SHAP se especifica como (Lundberg & Lee, 2017):

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \qquad (3.\ 10)$$

donde g es el modelo de explicación, $z' \in \{0,1\}^M$ es el vector de coalición o vector de características simplificadas, M es el tamaño máximo de coalición y $\phi_i \in \mathbb{R}$ es la atribución de característica para una característica i, lo que se conoce como valores de Shapley. En el vector de coalición, un valor de 1 significa que el valor de la característica está presente, y 0 que está ausente.

Los valores de SHAP son la única solución que satisface las propiedades de eficiencia, simetría y aditividad. Además, SHAP describe las siguientes tres propiedades deseables:

1) Precisión local.

Al aproximar el modelo original a una entrada específica x, la precisión local requiere que el modelo de explicación coincida al menos con la salida de la entrada simplificada x' (que corresponde a la entrada original x) (Lundberg & Lee, 2017):

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i \quad (3.\ 11)$$

Si definimos $\phi_0 = E_X(f(x))$ y configuramos todos los x'_i a 1, esta es la propiedad de eficiencia de Shapley. Únicamente con un nombre diferente y usando el vector de coalición (Lundberg & Lee, 2017):

$$(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i = E_X(f(X)) + \sum_{i=1}^{M} \phi_i \quad (3.\ 12)$$

2) Ausencia.

Si las entradas simplificadas representan la presencia de características, entonces la ausencia de estas entradas requiere que las características que faltan en la entrada original no tengan impacto (Lundberg & Lee, 2017):

$$x'_i = 0 \rightarrow \phi_i = 0 \quad (3.\ 13)$$

3) Coherencia.

La propiedad de coherencia especifica que si un modelo cambia de modo que la contribución marginal del valor de una característica aumenta o permanece igual (independientemente de otras características), el valor de Shapley también aumenta o permanece igual.

En este contexto, únicamente existe un modelo de explicación que satisface las propiedades 1, 2 y 3, y es (Lundberg & Lee, 2017):

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M-|z'|-1)!}{M!} [f_x(z') - f_x(z' \backslash i)] \quad (3.\ 14)$$

Donde $|z'|$ es el numero de entradas diferentes a 0 en *z'*, y $z' \subseteq x'$ representa todos los vectores de *z'* donde las entradas que no son nulas son un subconjunto de las entradas que no son nulas en *x'*.

En este trabajo, se ha utilizado la técnica SHAP para interpretar los resultados ofrecidos por el modelo propuesto. Esta técnica permitió detectar los píxeles de la imagen que más influyen en la clasificación de cada una de las clases.

Capítulo 4

Resultados

En este trabajo se han estudiado distintas arquitecturas CNN para la clasificación de la severidad de la RD en retinografías. Adicionalmente, se han utilizado las técnicas *data augmentation*, *dropout*, *transfer learning* y *fine-tuning* para optimizar las arquitecturas CNN.

Una vez que se ha descrito la metodología empleada en este trabajo, se presentan en este capítulo los principales resultados obtenidos en la clasificación automática de la RD. Se mostrará un análisis de la BD en términos de *t*-SNE. Se detallarán los resultados para cada una de las fases del método para tres escenarios diferentes: clasificación de la severidad de la RD, detección de la presencia de la RD y resultados de los casos derivables y no derivables de la RD. Asimismo, se describirán las métricas empleadas, teniendo en cuenta que no todos los autores utilizan las mismas.

4.1. Análisis de la base de datos

Considerando la BD de retinografías ATPOS-2019 descrita en el apartado 3.1, se procede a realizar un análisis de la misma. Para entender si las imágenes son separables en las respectivas clases que indican la gravedad de la RD, se ha empleado el algoritmo *t*-SNE. En la Figura 4.1 se muestra el resultado de aplicar *t*-SNE a la BD empleada en este trabajo. Cada color representa uno de los grados de severidad de la RD etiquetados por los expertos. El color azul claro representa los casos sanos, el color azul oscuro se corresponde con los casos de RDNP leve, el color verde hace referencia a los casos de RDNP moderada, el color amarillo indica una RDNP severa y el color rojo representa los casos de RD proliferativa. Como se puede observar, la clase 0, que indica ausencia de RD, está muy separada de las demás clases, mientras que la distinción entre el resto de clases resulta difícil.

Cabe destacar que los ejes de la gráfica no tienen ningún significado concreto. El algoritmo solo se centra en las distancias entre los puntos. Intenta colocar los puntos en un plano de modo que las distancias por pares entre ellos minimicen un cierto criterio. En este caso, el criterio sería la similitud entre los píxeles de las imágenes para intentar agrupar las más similares en determinados puntos en el espacio.

Para entender mejor cómo se realiza esta agrupación, en la Figura 4.2 se muestra un ejemplo de dos imágenes pertenecientes a la clase RDNP moderada. El algoritmo buscará para cada grupo de píxeles aquellas imágenes que comparten las mismas características en cada zona de la imagen. En la Figura 4.2 se observa que los píxeles situados en el interior de los recuadros negros comparten la presencia de exudados en la misma zona en ambas imágenes. Por tanto, el algoritmo agrupará estas dos imágenes en la misma parte del plano debido a su gran similitud. Este procedimiento se repite para cada grupo de píxeles de la imagen.

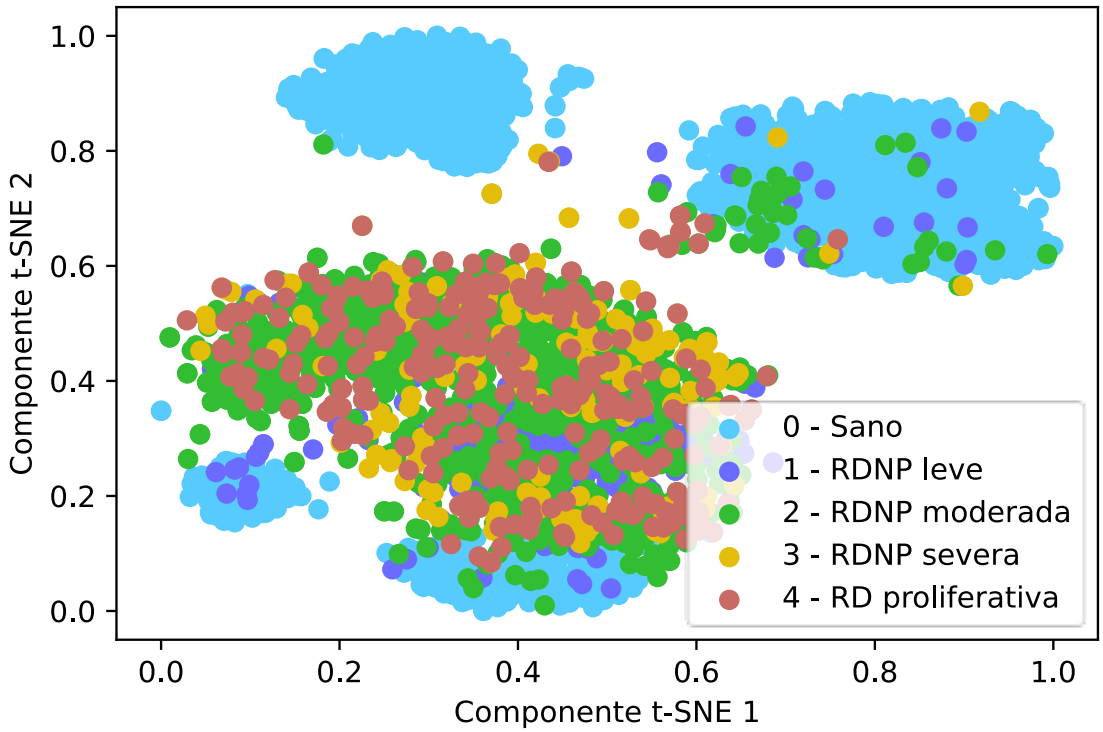


Figura 4.1 Visualización *t*-SNE de la BD de retinografías.

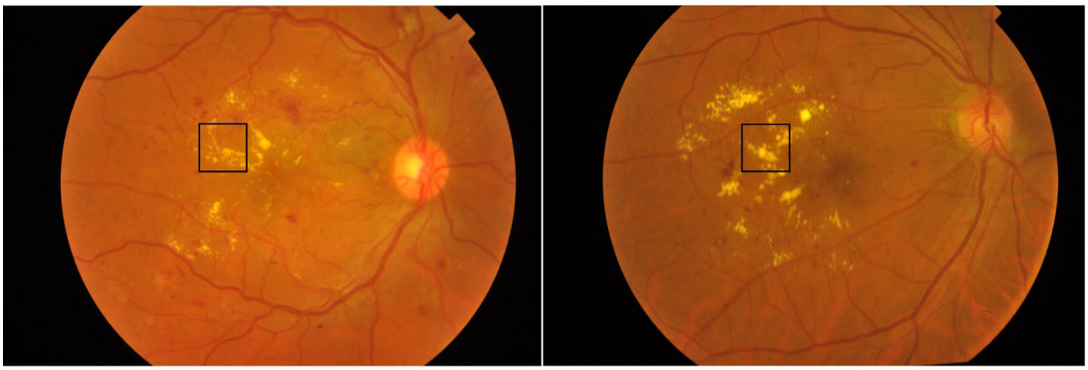


Figura 4.2 Ejemplo de imágenes pertenecientes a la clase RDNP moderada cuyos píxeles situados en el recuadro negro son similares.

4.2. Modo de evaluación

Un factor muy importante en la elaboración de un sistema de clasificación es la evaluación de los resultados. Es de gran importancia establecer medidas que nos permitan saber cómo se está comportando nuestro modelo. En la literatura existe una gran cantidad de métricas, aunque en este caso nos centraremos en algunas de las más comunes en la graduación de la severidad de la RD.

El problema analizado en este trabajo es un problema de clasificación, es decir la variable objetivo que predecimos solo puede tomar un conjunto de valores discretos. Además, se trata de un problema con cinco posibles clases: paciente sano, paciente con RDNP leve, paciente con RDNP moderada, paciente con RDNP severa y paciente con RD proliferativa. Al tratarse de un problema multiclase, el modelo CNN entrenado se ha evaluado en términos de pérdida, precisión o *accuracy*, coeficiente kappa, sensibilidad, especificidad, curva *Receiver Operating Characteristic* y matrices de confusión tanto en la fase de entrenamiento como de test (Majumder & Kehtarnavaz, 2021; Sikder *et ál.*, 2019).

4.2.1. Pérdida y precisión

Durante el entrenamiento de la red neuronal, la función de coste es clave para ajustar adecuadamente los pesos de una red neuronal y crear un modelo de aprendizaje automático que se ajuste mejor al problema en cuestión. En esta fase, se distinguen dos etapas: la propagación hacia delante y la propagación hacia atrás (Aggarwal, 2018). En la propagación hacia delante, la imagen se introduce en la capa de entrada y las neuronas aplican sobre ella distintas operaciones

matemáticas. El resultado se envía a la capa de salida generando la predicción final, que, en el caso de los problemas de clasificación, indican la probabilidad o confianza de pertenecer a las posibles etiquetas o clases (Aggarwal, 2018; Ho & Wookey, 2019). A continuación, en la propagación hacia atrás, estas probabilidades se comparan con las etiquetas reales y la función de coste calcula una penalización por cualquier desviación entre la etiqueta real y la predicción de la red neuronal, siendo en este caso el grado de RD que presenta la retinografía. Si el error cometido es grande, los pesos de la red se modifican y se actualizan hacia atrás, de manera que el error decrece a medida que el entrenamiento avanza. Este proceso se repite hasta conseguir un modelo óptimo que permite la clasificación correcta de las imágenes de fondo de ojo en sus respectivas etapas de la RD (Ho & Wookey, 2019).

La entropía cruzada es la función de pérdida de facto en las tareas de clasificación modernas que implican distinguir cientos o incluso miles de clases, y, para el caso de clasificación multiclase, como el de este TFM, esta pérdida viene dada por (Ho & Wookey, 2019):

$$Pérdida_{entropía\,cruzada} = -\frac{1}{M}\sum_{k=1}^{K}\sum_{m=1}^{M} y_m^k \times \log\left(h_\theta(x_m, k)\right) \qquad (4.\ 1)$$

, donde M es el número de imágenes que hay durante el entrenamiento, K es el número de clases, y_m^kes la etiqueta de destino para el ejemplo de entrenamiento m y la clase k, x es la entrada del ejemplo de entrenamiento m, y h_θ es el modelo con los pesos θ de la red neuronal.

Respecto a la precisión, la métrica utilizada ha sido la precisión categórica o *categorical accuracy*, la cual calcula el porcentaje de valores pronosticados por la CNN que coinciden con los valores reales. Se calcula como (Grandini *et ál.*, 2020):

$$Precisión = \frac{VP + VN}{VP + VN + FP + FN} \qquad (4.\ 2)$$

, donde VP es el número de verdaderos positivos, VN el número de verdaderos negativos, FP el de falsos positivos y FN el número de falsos negativos.

4.2.2. Matriz de confusión

La matriz de confusión es una tabla cruzada que registra la concordancia en la clasificación entre dos evaluadores. En este caso, se utilizará para cruzar la clasificación real evaluada por el experto y la clasificación predicha por el método propuesto. Las columnas representan la predicción del modelo mientras que las filas muestran la verdadera clasificación. Un ejemplo genérico de matriz de confusión se puede observar en la Figura 4.3.

Las clases se enumeran en el mismo orden en las filas que en las columnas, por lo que los elementos correctamente clasificados se ubican en la diagonal principal de arriba a abajo a la derecha y corresponden al número de veces que los dos evaluadores están de acuerdo (Grandini *et ál.*, 2020).

4.2.3. Coeficiente kappa de Cohen

El coeficiente kappa de Cohen se basa en la idea de medir la concordancia entre las etiquetas predichas y verdaderas, que se consideran dos variables categóricas aleatorias (Ranganathan *et ál.*, 2017). Es posible comparar dos variables categóricas construyendo la matriz de confusión y calculando las distribuciones de filas marginales y columnas marginales, como se muestra en la Figura 4.3. En esta figura, a indica la clase real, mientras que p representa la clase predicha por el modelo.

		Predicción del modelo					
		p = 1	…	p = h	…	p = K	Total
Clasificación real	a = 1	n_{11}	…	n_{1h}	…	n_{1K}	n_{1T}
	…	…	…	…	…	…	…
	a = v	n_{v1}	…	n_{vh}	…	n_{vK}	n_{vT}
	…	…	…	…	…	…	…
	a = K	n_{K1}	…	n_{Kh}	…	n_{KK}	n_{KT}
	Total	n_{T1}	…	n_{Th}	…	n_{TK}	$\boldsymbol{N}$

Figura 4.3 Matriz de confusión general para una distribución categórica.

En particular, dos distribuciones del mismo carácter son independientes si asumen las mismas frecuencias relativas en el mismo modelo (Ranganathan *et ál.*, 2017):

$$\frac{n_{vh}}{n_{Th}} = \frac{n_{vT}}{N} \tag{4. 3}$$

Además, la predicción del modelo y la clasificación real son variables independientes en la distribución si se cumple (Ranganathan *et ál.*, 2017):

$$n^*_{vh} = \frac{n_{Th} \times n_{vT}}{N} \tag{4. 4}$$

, donde n^*_{vh} representa una frecuencia relativa que esperamos encontrar si las distribuciones categóricas son independientes.

Dada esta definición de independencia entre variables categóricas, se pueden emplear los indicadores kappa de Cohen como valores de calificación de la dependencia (o independencia) entre la predicción del modelo y la clasificación real.

En el caso de clasificación multiclase, el cálculo del coeficiente kappa de Cohen se define como (Ranganathan *et ál.*, 2017):

$$K = \frac{c \times s - \sum_k^K p_k \times t_k}{s^2 - \sum_k^K p_k \times t_k} \tag{4. 5}$$

, donde:

$c = \sum_k^K C_{kk}$ es el número de elementos predichos correctamente

$s = \sum_i^K \sum_j^K C_{ij}$ es el número total de elementos

$p_k = \sum_i^K C_{ki}$ es el número de veces que se predijo la clase k (total de la columna)

$t_k = \sum_i^K C_{ik}$ es el número de veces que ocurre realmente la clase k (total de filas)

4.2.4. Curva ROC

La curva *Receiver Operating Characteristic* (ROC) permite visualizar el rendimiento de un modelo representando la sensibilidad frente a la especificidad. La sensibilidad se refiere a la capacidad de identificar correctamente las entradas que pertenecen a la clase positiva. La especificidad se refiere a la capacidad de identificar correctamente las entradas que pertenecen a la clase negativa. Se pueden expresar estos dos parámetros como (Fawcett, 2006):

$$Sensibilidad = \frac{VP}{VP + FN} \tag{4. 6}$$

$$Especificidad = \frac{VN}{VN + FP} \tag{4. 7}$$

donde VP es el número de verdaderos positivos, VN el número de verdaderos negativos, FP el de falsos positivos y FN el número de falsos negativos.

La información de la curva ROC se puede sintetizar en un único parámetro, el área bajo la curva ROC (*Area Under Curve*, AUC) (Briggs *et ál.*, 2008). En este contexto, un valor de AUC de 1 significa que el modelo puede distinguir con precisión entre las distintas clases el 100% del tiempo. Una puntuación de 0.5 significa que el modelo no puede discriminar entre las clases, y está tratando de adivinar la clase (Wandishin & Mullen, 2009).

En el caso de clasificación multiclase, como ocurre en este trabajo, las curvas ROC se pueden trazar con la metodología de usar una clase frente al resto. Esto significa que obtendremos una curva para cada clase, y, en consecuencia, un parámetro AUC para cada clase individual (Wandishin & Mullen, 2009).

4.3. Medida de resultados

4.3.1. Conjuntos de entrenamiento, validación y test

El conjunto de imágenes de la BD utilizada se ha dividido en tres subconjuntos: entrenamiento, validación y test. El conjunto de entrenamiento se utilizó como datos de aprendizaje para ajustar los parámetros internos de la red.

El conjunto de validación permitió evaluar el rendimiento del modelo a cada época de entrenamiento, lo que sirvió para verificar el correcto aprendizaje y evitar el sobreentrenamiento. Por último, el conjunto de test, formado por imágenes independientes, se utilizó para obtener una evaluación del sistema final con datos nuevos y comprobar el funcionamiento del sistema de predicción implementado.

Las imágenes se han dividido de manera aleatoria para formar un conjunto de entrenamiento, uno de validación y otro de test. Para la división de la BD se ha elegido una proporción 80:10:10. Esto da como resultado un conjunto de entrenamiento formado por 2930 imágenes (80% de 3662) mientras que los conjuntos de validación y test están formados por 366 imágenes (10% de 3662). Dichos datos se muestran en la Tabla 4.1.

4.3.2. Fase de entrenamiento

Los objetivos de la fase de entrenamiento han sido: determinar qué arquitectura CNN permite obtener los mejores resultados, y comprobar los resultados que se obtienen en los conjuntos de entrenamiento y validación. El procedimiento que se ha seguido en cada una de las fases de entrenamiento ha sido el siguiente:

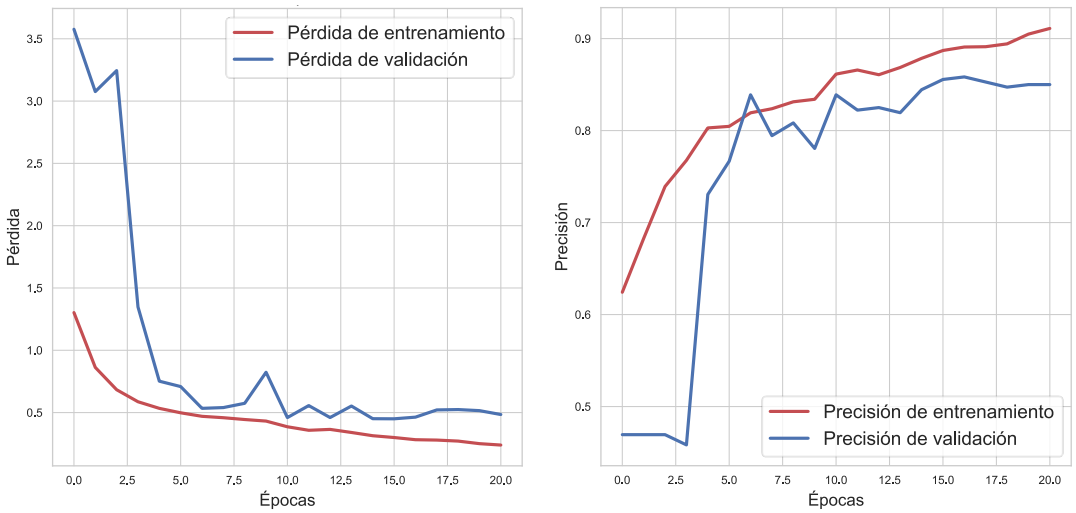
1. Determinar la arquitectura más adecuada entre ResNet-50, Inception-V3, DenseNet-201 y MobileNet-V2. Para ello, se ha calculado el coeficiente kappa, sensibilidad, especificidad, AUC y precisión sobre los conjuntos de entrenamiento y de validación para todas las arquitecturas. Estas son las principales métricas que se han empleado en estudios anteriores (Alyoubi *et ál.*, 2021; Bodapati *et ál.*, 2020; Dekhil *et ál.*, 2019; Pham *et ál.*, 2020).

Conjunto de imágenes	**Número de imágenes**
Conjunto de entrenamiento	2930
Conjunto de validación	366
Conjunto de test	366
Total	3662

Tabla 4.1 Separación de las imágenes para la tarea de clasificación de la severidad de la RD en conjuntos de entrenamiento, validación y test.

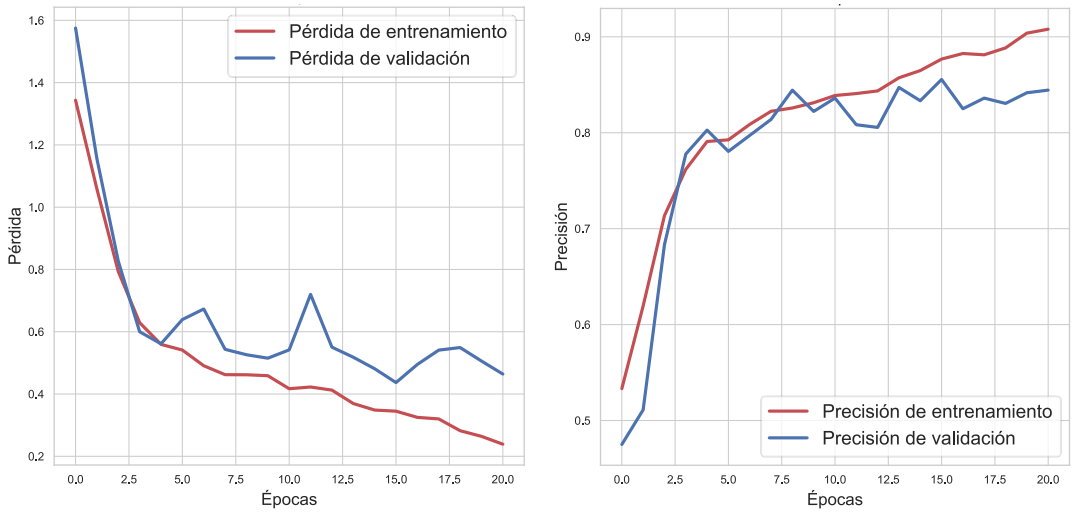
2. Una vez determinada la arquitectura óptima se calculan también las curvas ROC y la matriz de confusión para la arquitectura escogida. Esto ofrece una visión más detallada del funcionamiento del modelo de manera que se puede comprobar el correcto aprendizaje de la red.

En primer lugar, la evolución de la pérdida y precisión para cada época de entrenamiento y para todas las redes se muestran en las Figuras 4.4 - 4.7. En todos los casos se puede apreciar que la pérdida tiende a decrecer a medida que avanzan las épocas de entrenamiento, mientras que la precisión tiende a aumentar. Para todas las arquitecturas, se considera que el modelo no está sobreentrenado, ya que los resultados obtenidos en el grupo de validación mejoran en la línea de los del grupo de entrenamiento. El concepto sobreentrenamiento implica que el modelo se ajuste demasiado al conjunto de entrenamiento y no sea capaz de generalizar al clasificar otras imágenes independientes, como serían los conjuntos de validación y de test. Este es un factor muy importante a tener en cuenta a la hora de desarrollar un método de clasificación automático de RD, puesto que el objetivo es que el sistema sea capaz de clasificar imágenes de fondo de ojo con características procedentes de diferentes pacientes.



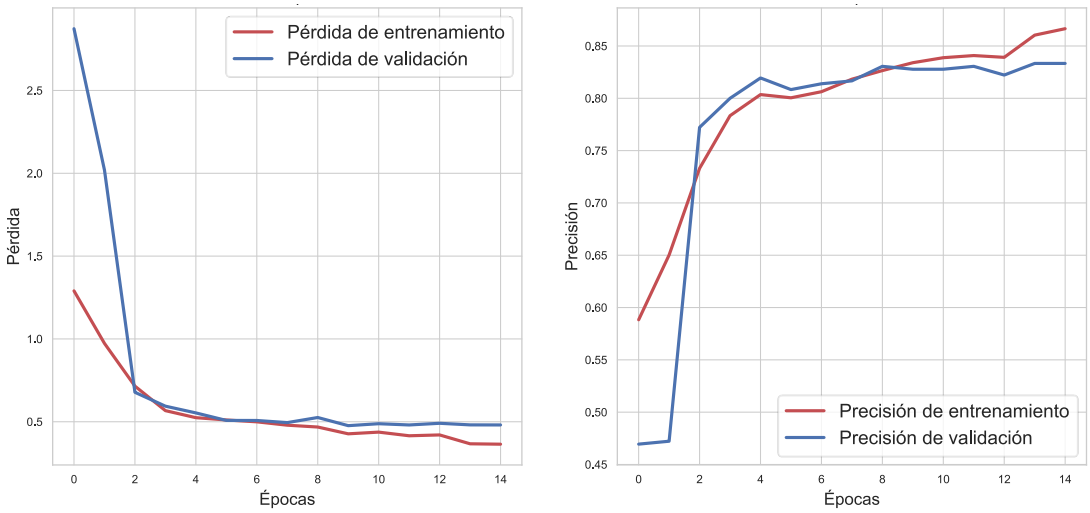
(a) **(b)**

Figura 4.4 Arquitectura ResNet-50. **(a)** Evolución de la pérdida en función de las épocas de entrenamiento, para los conjuntos de entrenamiento y validación. **(b)** Evolución de la precisión en función de las épocas de entrenamiento, para los conjuntos de entrenamiento y validación.



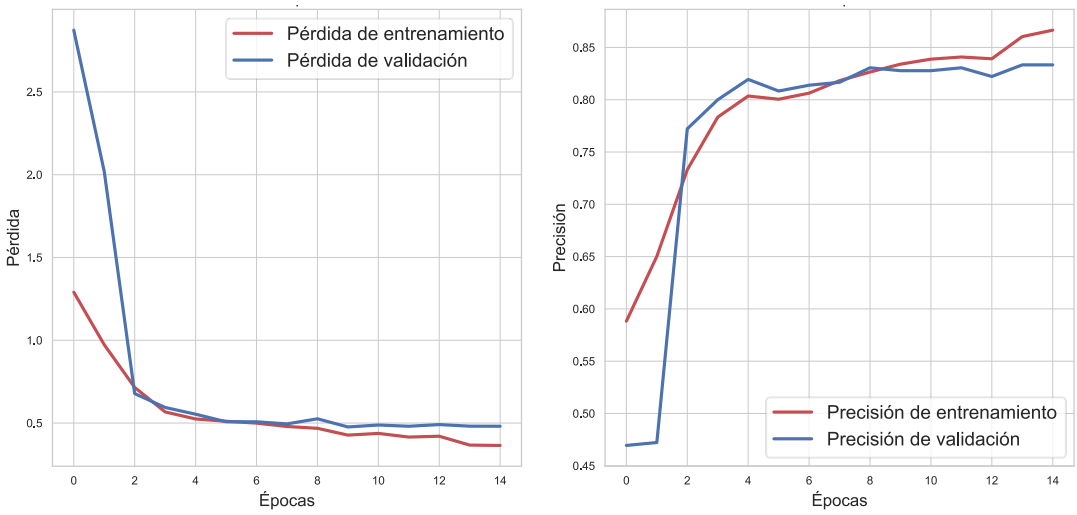
(a) **(b)**

Figura 4.5 Arquitectura Inception-V3. **(a)** Evolución de la pérdida en función de las épocas de entrenamiento, para los conjuntos de entrenamiento y validación. **(b)** Evolución de la precisión en función de las épocas de entrenamiento, para los conjuntos de entrenamiento y validación.



(a) **(b)**

Figura 4.6 Arquitectura DenseNet-201. **(a)** Evolución de la pérdida en función de las épocas de entrenamiento, para los conjuntos de entrenamiento y validación. **(b)** Evolución de la precisión en función de las épocas de entrenamiento, para los conjuntos de entrenamiento y validación.



(a) **(b)**

Figura 4.7 Arquitectura MobileNet-V2. **(a)** Evolución de la pérdida en función de las épocas de entrenamiento, para los conjuntos de entrenamiento y validación. **(b)** Evolución de la precisión en función de las épocas de entrenamiento, para los conjuntos de entrenamiento y validación.

4.3.2.1 Resultados sobre la clasificación de la severidad de la RD

El primero de los escenarios planteados se corresponde a la clasificación de la severidad de la RD en cinco clases: ausencia de RD, RDNP leve, RDNP moderada, RDNP severa y RD proliferativa. En las Tablas 4.2 y 4.3 se muestran los resultados para todas las arquitecturas CNN sobre los conjuntos de entrenamiento y validación, respectivamente. Se puede apreciar que se han obtenido resultados con un elevado coeficiente kappa, especificidad, AUC y precisión para todas las arquitecturas. De hecho, no hay mucha diferencia entre los resultados de las

Arquitectura CNN	Coeficiente kappa	Sensibilidad	Especificidad	AUC	Precisión
ResNet-50	**0.952**	**80.60%**	**97.80%**	**0.98**	**96.20%**
Inception-V3	0.947	79.20%	97.80%	0.98	96%
DenseNet-201	0.892	70.80%	96.60%	0.96	94.80%
MobileNet-V2	0.918	71.20%	96.60%	0.97	94.60%

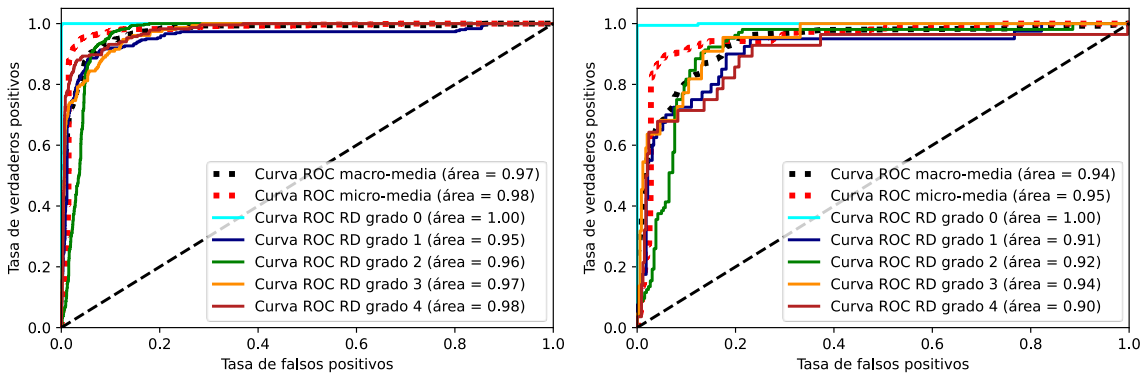
Tabla 4.2 Resultados sobre el conjunto de entrenamiento para la clasificación de la severidad de la RD.

Arquitectura CNN	Coeficiente kappa	Sensibilidad	Especificidad	AUC	Precisión
ResNet-50	**0.904**	**70.20%**	**96%**	**0.95**	**94.20%**
Inception-V3	0.898	64%	95.80%	0.95	93.40%
DenseNet-201	0.898	63.80%	95.80%	0.95	93.20%
MobileNet-V2	0.889	63.20%	95.20%	0.95	93.20%

Tabla 4.3 Resultados sobre el conjunto de validación para la clasificación de la severidad de la RD.

distintas arquitecturas. Sin embargo, la arquitectura CNN que presenta un mayor valor en todas las métricas es ResNet-50. Por lo tanto, se considera que esta sería la arquitectura óptima para el problema.

Una vez escogida la CNN óptima, se muestran el resto de los resultados únicamente para esta arquitectura. El análisis ROC del aprendizaje del modelo en los conjuntos de entrenamiento y validación se presenta en la Figura 4.8. Se puede observar cómo, para los pacientes sanos, el AUC es 1 en ambos casos. Esto significa que, en la fase de entrenamiento, el modelo es capaz de clasificar correctamente la totalidad de los sujetos sin RD. Por otra parte, la clase RD proliferativa es la que menor AUC obtiene, lo que significa que la clasificación de esta clase es menos precisa. En esta figura se incluyen los micro-medios y los macro-medios. El macro-medio calcula la métrica de forma independiente para cada clase y después realiza la media, tratando todas las clases por igual. Por otra parte, el micro-medio tiene en cuenta el número de muestras de cada clase para

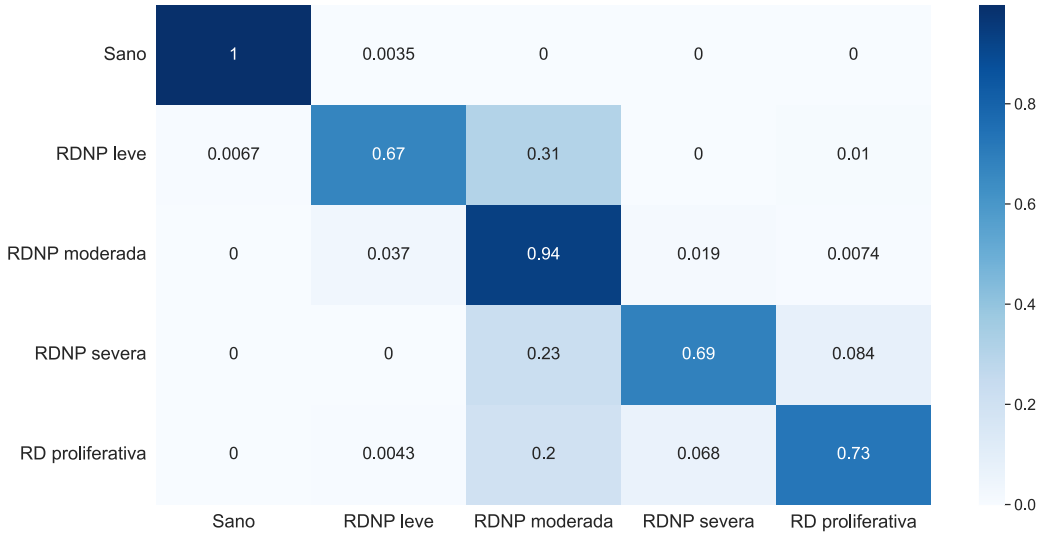


(a) **(b)**

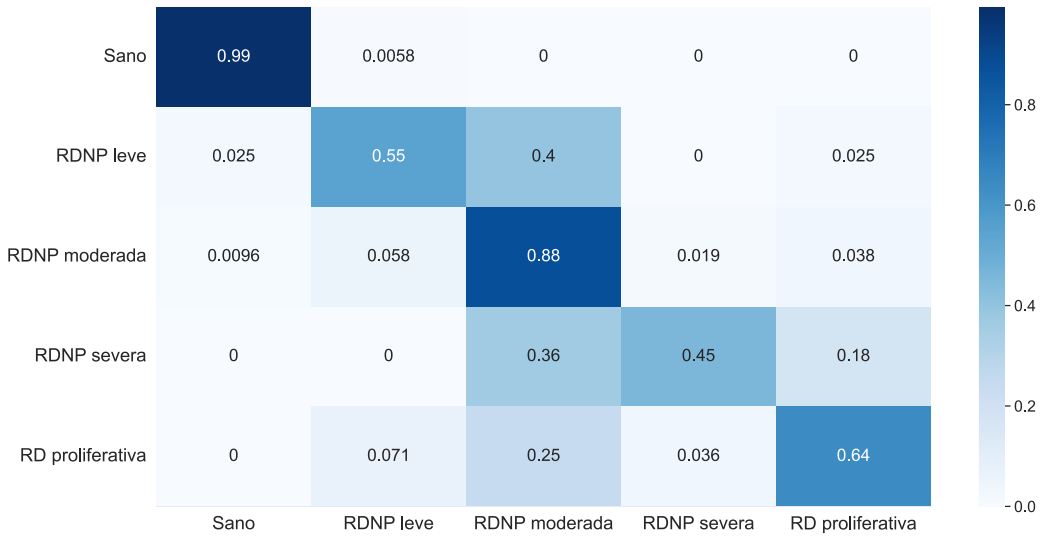
Figura 4.8 (a) Curva ROC de cada grado de severidad de la RD frente al resto para el conjunto de entrenamiento. **(b)** Curva ROC de cada grado de severidad de la RD frente al resto para el conjunto de validación.

calcular la métrica media (Gowda *et ál.*, 2021). En el caso de un problema de clasificación multiclase con desequilibrio o desbalanceo de clases como ocurre en este caso, es más adecuado calcular el micro-medio (Gowda *et ál.*, 2021).

Por último, la matriz de confusión normalizada para los mismos conjuntos se muestra en la Figura 4.9. Se observa cómo la diagonal obtiene los valores más altos, lo que indica que los valores predichos por la red coinciden en gran medida con las clases reales etiquetadas por los oftalmólogos.



(a)



(b)

Figura 4.9 (a) Matriz de confusión multiclase normalizada sobre el conjunto de entrenamiento. **(b)** Matriz de confusión multiclase normalizada sobre el conjunto de validación.

4.3.2.2 Resultados sobre la detección de la presencia de la RD

En este segundo escenario, se lleva a cabo un análisis de la detección de la presencia de la RD. Para ello, se consideran los casos de ausencia de RD frente a los casos de RD, es decir, RDNP leve, RDNP moderada, RDNP severa y RD proliferativa.

En las Tablas 4.4 y 4.5 se muestran los resultados sobre los conjuntos de entrenamiento y validación, respectivamente. Se puede apreciar que se han obtenido resultados con un elevado coeficiente kappa, sensibilidad, especificidad, AUC y precisión para todas las arquitecturas. Aunque en el conjunto de entrenamiento la arquitectura CNN que presenta mayores métricas es Inception-V3, ResNet-50 es la que mejor responde sobre el conjunto de validación. Se puede observar que la arquitectura ResNet-50 presenta un mayor valor de coeficiente kappa, especificidad, AUC y precisión. Por lo tanto, de nuevo se considera que esta sería la arquitectura óptima para la detección de la presencia de la RD.

Una vez escogida la CNN óptima, se muestran el resto de los resultados únicamente para esta arquitectura. El análisis ROC del aprendizaje del modelo en los conjuntos de entrenamiento y validación se presenta en la Figura 4.10.

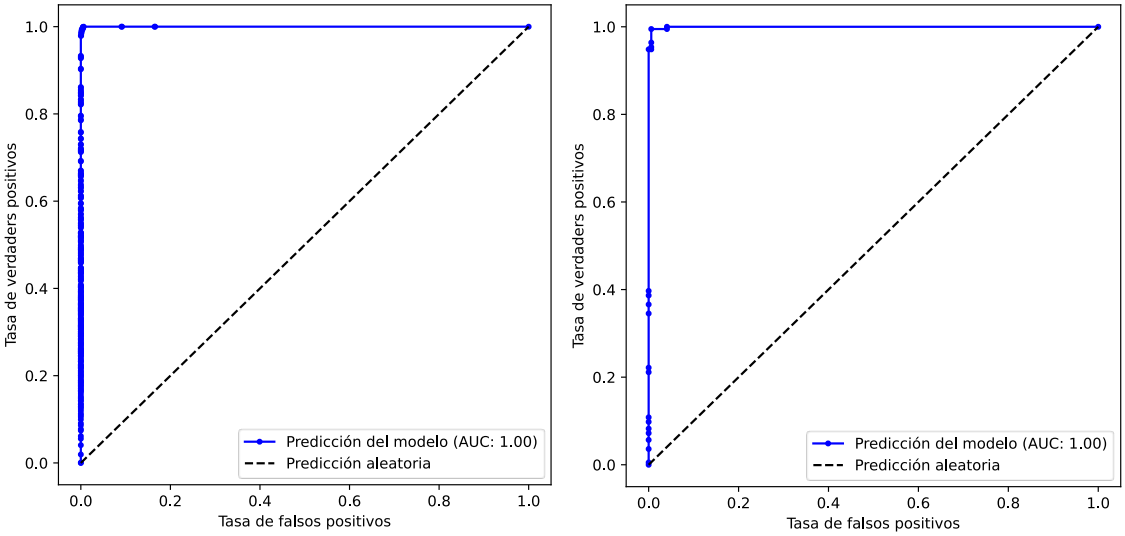
Arquitectura CNN	Coeficiente kappa	Sensibilidad	Especificidad	AUC	Precisión
ResNet-50	0.988	98.54%	100%	1	99.28%
Inception-V3	**0.994**	**99.37%**	**100%**	**1**	**99.69%**
DenseNet-201	0.982	98.53%	99.60%	1	99.08%
MobileNet-V2	0.987	98.95%	99.73%	1	99.35%

Tabla 4.4 Resultados sobre el conjunto de entrenamiento para la detección de la presencia de la RD.

Arquitectura CNN	Coeficiente kappa	Sensibilidad	Especificidad	AUC	Precisión
ResNet-50	**0.989**	98.83%	**100%**	**1**	**99.45%**
Inception-V3	0.988	99.41%	99.48%	1	99.45%
DenseNet-201	0.984	99.41%	98.96%	1	99.18%
MobileNet-V2	0.988	99.41%	99.48%	1	99.45%

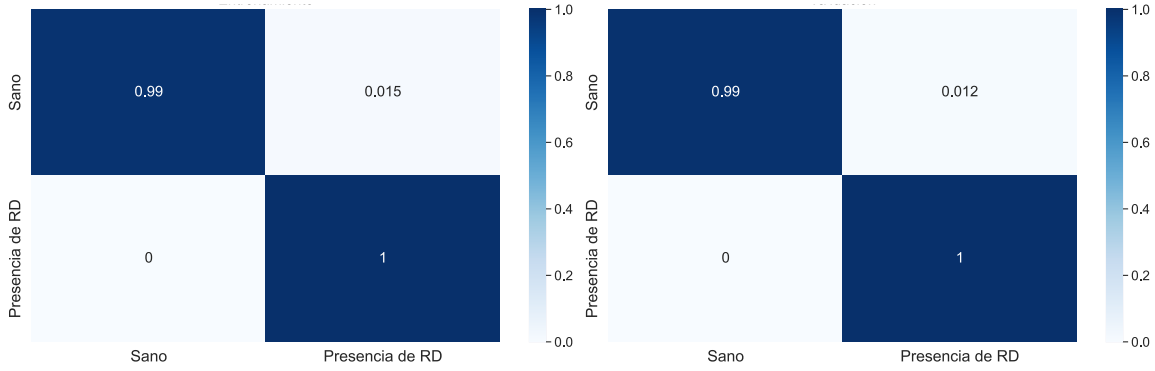
Tabla 4.5 Resultados sobre el conjunto de validación para la detección de la presencia de la RD.

Las matrices de confusión normalizadas para los mismos conjuntos se muestran en la Figura 4.11. Se observa cómo la diagonal obtiene los valores más altos, lo que indica que los valores predichos por la red coinciden en gran medida con las clases reales etiquetadas por los oftalmólogos.



(a) **(b)**

Figura 4.10 **(a)** Curva ROC de la detección de la presencia de RD en el conjunto de entrenamiento. **(b)** Curva ROC de la detección de la presencia de RD en el conjunto de validación.



(a) **(b)**

Figura 4.11 **(a)** Matriz de confusión normalizada de la detección de la presencia de RD en el conjunto de entrenamiento. **(b)** Matriz de confusión normalizada de la detección de la presencia de RD en el conjunto de validación.

4.3.2.3 Resultados de los casos derivables y no derivables de la RD

En este escenario se considera la división en casos derivables frente a casos no derivables. Se ha considerado que los casos no derivables pertenecen a los casos de ausencia de RD y RDNP leve, mientras que los casos derivables son aquellos que presentan RDNP moderada, RDNP severa y RD proliferativa (Gwenolé Quellec *et ál.*, 2017).

En las Tablas 4.6 y 4.7 se muestran los resultados sobre los conjuntos de entrenamiento y validación, respectivamente. Se han obtenido resultados con una elevada especificidad, AUC y precisión para todas las arquitecturas. Realmente, no hay mucha diferencia entre los resultados de las distintas arquitecturas. En el conjunto de entrenamiento, la arquitectura que presenta mayores valores de todas las métricas en Inception-V3. Sin embargo, la arquitectura CNN que presenta las métricas más altas de coeficiente kappa, sensibilidad, especificidad y precisión sobre el conjunto de validación es ResNet-50. Por lo tanto, esta sería la arquitectura óptima en este escenario.

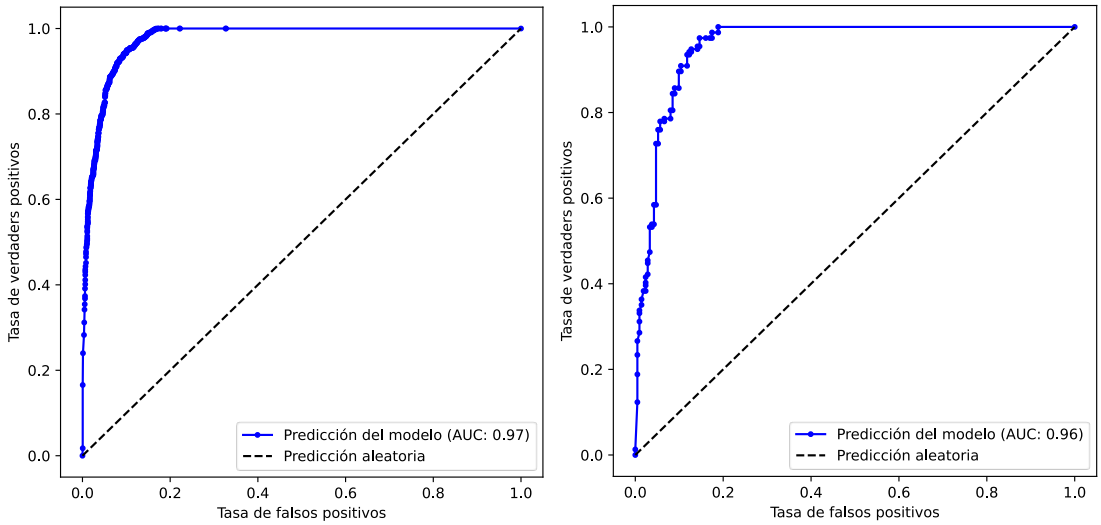
Arquitectura CNN	Coeficiente kappa	Sensibilidad	Especificidad	AUC	Precisión
ResNet-50	0.782	81.49%	100%	0.97	89.04%
Inception-V3	**0.790**	**82.18%**	**100%**	**0.98**	**89.45%**
DenseNet-201	0.784	81.72%	99.83%	0.97	89.11%
MobileNet-V2	0.787	82%	99.92%	0.97	89.32%

Tabla 4.6 Resultados sobre el conjunto de entrenamiento de los casos derivables y no derivables de la RD.

Arquitectura CNN	Coeficiente kappa	Sensibilidad	Especificidad	AUC	Precisión
ResNet-50	**0.782**	**80.66%**	**100%**	0.96	**88.52%**
Inception-V3	0.773	80.66%	99.35%	**0.97**	88.52%
DenseNet-201	0.778	81.13%	99.35%	0.96	88.52%
MobileNet-V2	0.773	80.66%	99.35%	0.95	88.52%

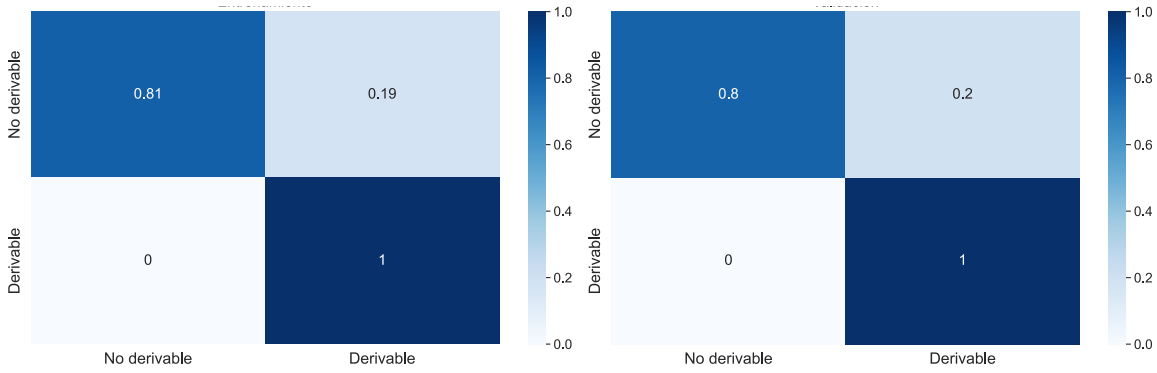
Tabla 4.7 Resultados sobre el conjunto de validación de los casos derivables y no derivables de la RD.

Una vez escogida la CNN óptima, se muestran el resto de los resultados únicamente para esta arquitectura. El análisis ROC del aprendizaje del modelo en los conjuntos de entrenamiento y validación se presenta en la Figura 4.12. Las matrices de confusión normalizadas para los mismos conjuntos se muestran en la Figura 4.13. Se observa cómo la diagonal obtiene los valores más altos, lo que indica que los valores predichos por la red coinciden en gran medida con la clasificación realizada por los oftalmólogos.



(a) **(b)**

Figura 4.12 **(a)** Curva ROC de los casos derivables y no derivables de RD en el conjunto de entrenamiento. **(b)** Curva ROC de los casos derivables y no derivables presencia de RD en el conjunto de validación.



(a) **(b)**

Figura 4.13 **(a)** Matriz de confusión normalizada de los casos derivables y no derivables de RD en el conjunto de entrenamiento. **(b)** Matriz de confusión normalizada de los casos derivables y no derivables presencia de RD en el conjunto de validación.

4.3.3. Fase de test

En la fase de test, se aplican los parámetros óptimos obtenidos en la fase de entrenamiento sobre el conjunto de imágenes de test. Así se puede comprobar la efectividad del algoritmo desarrollado en un conjunto de imágenes distinto al conjunto en que se realizó el entrenamiento. De nuevo, se muestran los resultados para los tres escenarios planteados utilizando la arquitectura CNN ResNet-50, la cual se ha seleccionado como óptima en la fase de entrenamiento.

4.3.3.1 Resultados sobre la clasificación de la severidad de la RD

Los resultados obtenidos sobre la clasificación de la severidad de la RD se dividen en dos tablas. En la Tabla 4.8 se muestran los valores de coeficiente kappa, sensibilidad, especificidad, AUC y precisión obtenidos sobre el conjunto de test. En la Tabla 4.9 se pueden observar los valores de sensibilidad, especificidad, AUC y precisión para cada clase en concreto. A continuación, en la Figura 4.14 se muestra la curva ROC obtenida sobre el conjunto de imágenes de test. Finalmente, en la Figura 4.15 se presenta la matriz de confusión normalizada que se ha obtenido en la fase de test. Por lo general, se puede apreciar que la diagonal principal presenta los valores más altos, lo que da una idea de que el método desarrollado generaliza bien a otro tipo de imágenes no empleadas en la fase de entrenamiento.

Arquitectura CNN	**Coeficiente kappa**	**Sensibilidad**	**Especificidad**	**AUC**	**Precisión**
ResNet-50	**0.919**	66.82%	**96.31%**	0.93	94.10%

Tabla 4.8 Métricas obtenidas sobre el conjunto de test en la clasificación de la severidad de la RD.

Métrica	**Ausencia de RD**	**RDNP leve**	**RDNP moderada**	**RDNP severa**	**RD proliferativa**
Sensibilidad	97.99%	50%	87.36%	41.18%	57.58%
Especificidad	98.80%	97.02%	88.89%	97.13%	99.70%
AUC	0.99	0.88	0.94	0.89	0.93
Precisión	98.36%	93.17%	88.52%	94.54%	95.90%

Tabla 4.9 Métricas obtenidas para cada clase sobre el conjunto de test en la clasificación de la severidad de la RD.

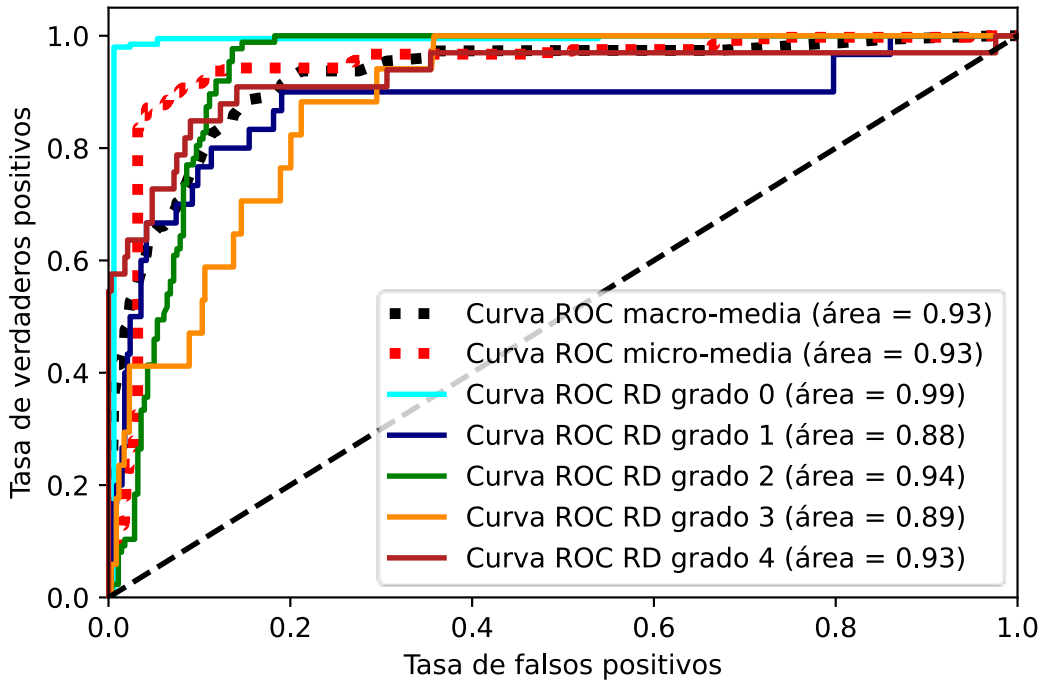


Figura 4.14 Curvas ROC de cada grado de severidad de la RD frente al resto para el conjunto de test.

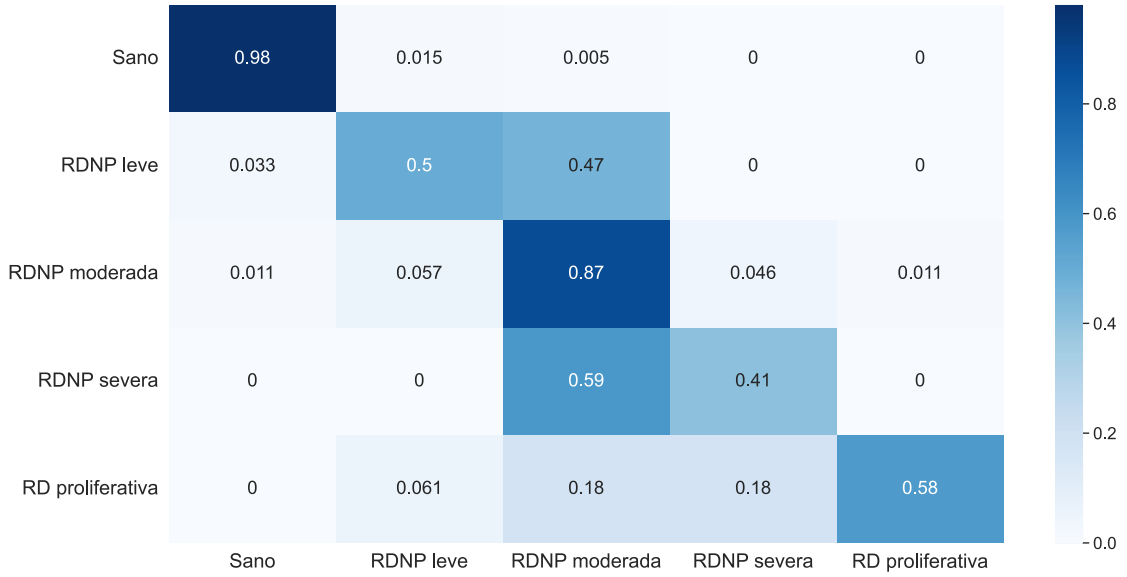


Figura 4.15 Matriz de confusión multiclase normalizada sobre el conjunto de test.

4.3.3.2 Resultados sobre la detección de la presencia de la RD

Los resultados obtenidos en este escenario se resumen en la Tabla 4.10 donde se muestran los valores de coeficiente kappa, sensibilidad, especificidad, AUC y precisión obtenidos sobre el conjunto de test. A continuación, en la Figura 4.16 se muestra la curva ROC obtenida sobre el conjunto de imágenes de test. Finalmente, en la Figura 4.17 se presenta la matriz de confusión que se ha obtenido en la fase de test. Se observa que la diagonal principal presenta los valores más altos, lo que significa que el método desarrollado generaliza bien a otro tipo de imágenes no empleadas en la fase de entrenamiento.

Arquitectura CNN	Coeficiente kappa	Sensibilidad	Especificidad	AUC	Precisión
ResNet-50	0.978	97.99%	100%	1	98.91%

Tabla 4.10 Métricas obtenidas sobre el conjunto de test en la detección de la presencia de la RD.

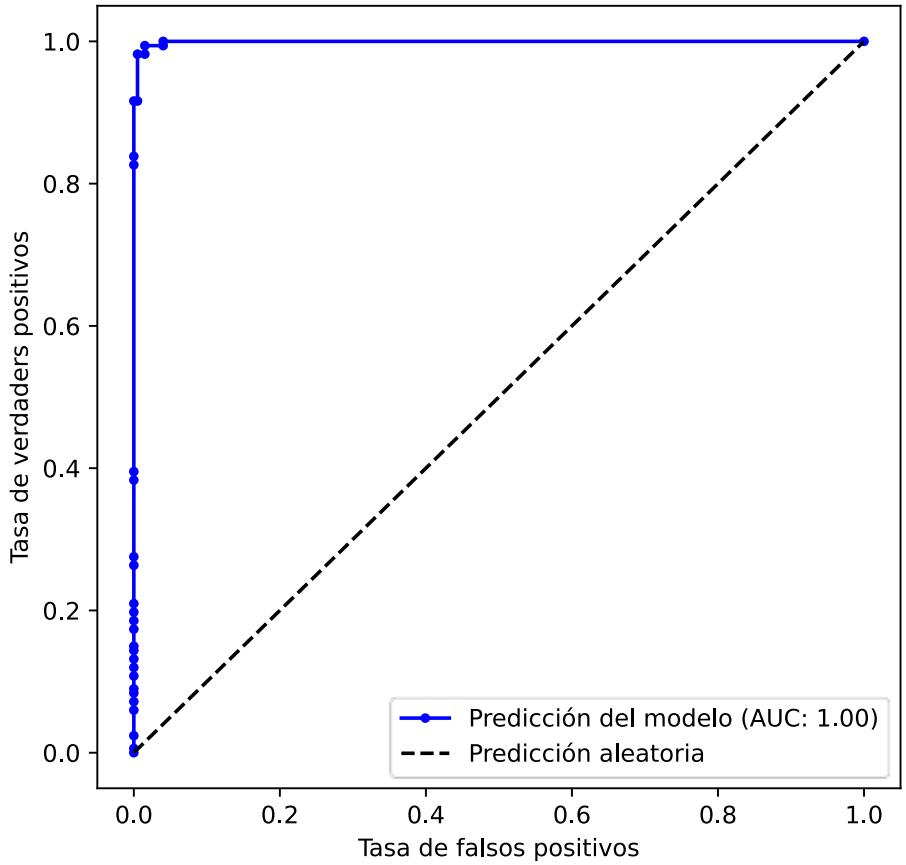


Figura 4.16 Curva ROC de la detección de la presencia de la RD sobre el conjunto de test.

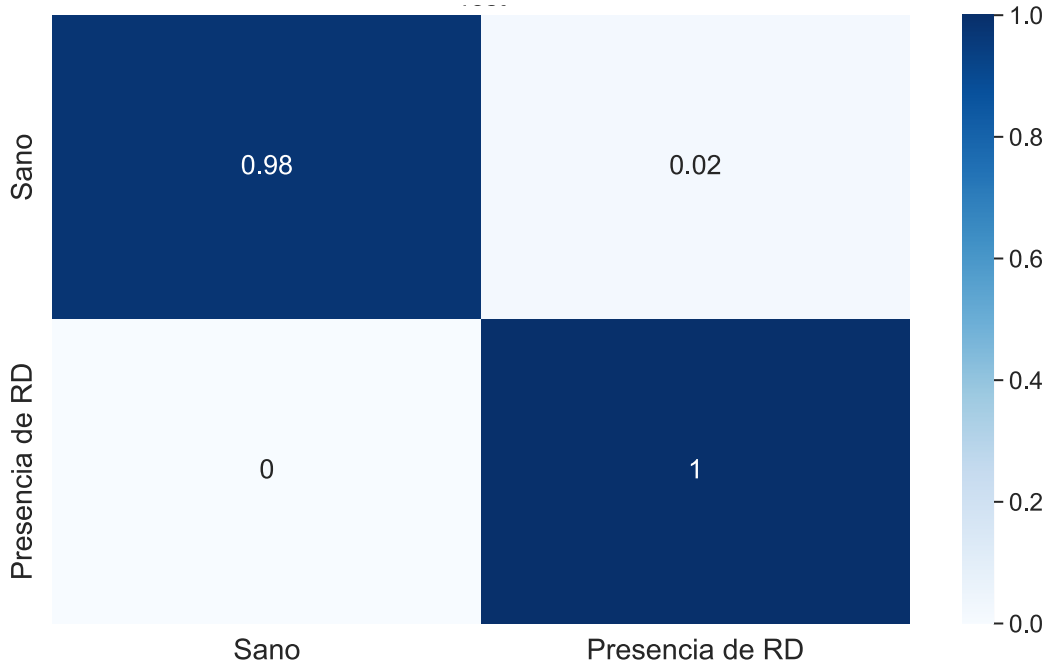


Figura 4.17 Matriz de confusión de la detección de la presencia de la RD sobre el conjunto de test.

4.3.3.3 Resultados de los casos derivables y no derivables de la RD

En la Tabla 4.11 se muestran los valores de coeficiente kappa, sensibilidad, especificidad, AUC y precisión sobre el conjunto de test para la división entre casos derivables y no derivables. Seguidamente, en la Figura 4.18 se muestra la curva ROC obtenida sobre el conjunto de imágenes de test. Por último, en la Figura 4.19 se presenta la matriz de confusión normalizada que se ha obtenido en la fase de test. La diagonal principal presenta los valores más altos, lo que da una idea de que el método desarrollado generaliza bien a otro tipo de imágenes no empleadas en la fase de entrenamiento.

Arquitectura CNN	**Coeficiente kappa**	**Sensibilidad**	**Especificidad**	**AUC**	**Precisión**
ResNet-50	**0.811**	85.15%	**100%**	0.97	90.71%

Tabla 4.11 Métricas obtenidas sobre el conjunto de test sobre los casos derivables y no derivables de la RD.

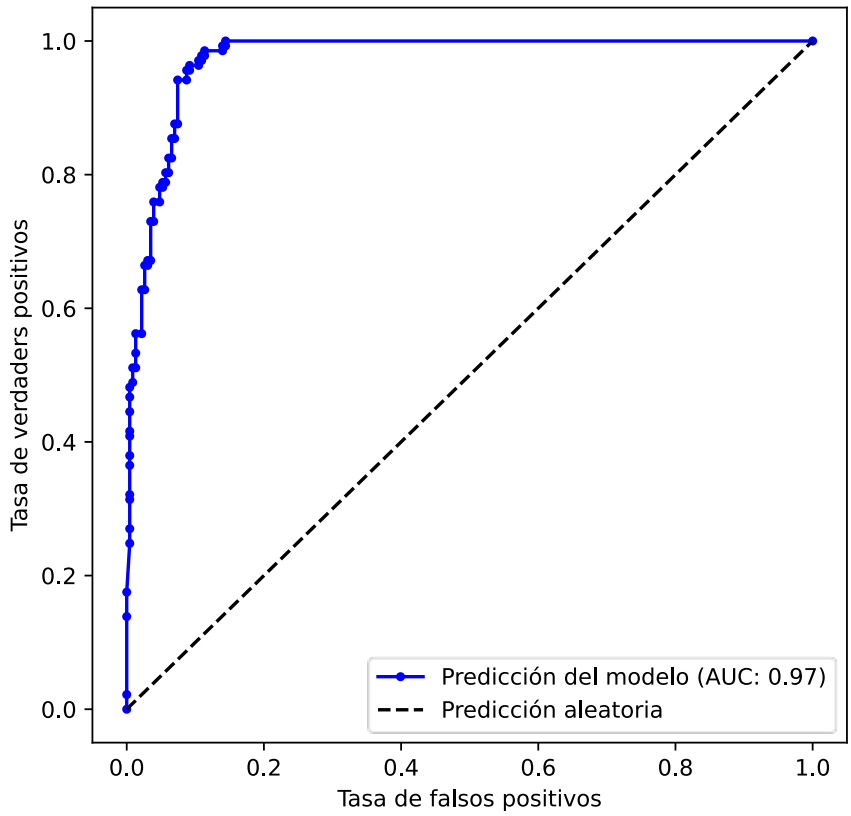


Figura 4.18 Curva ROC de los casos derivables y no derivables de la RD sobre el conjunto de test.

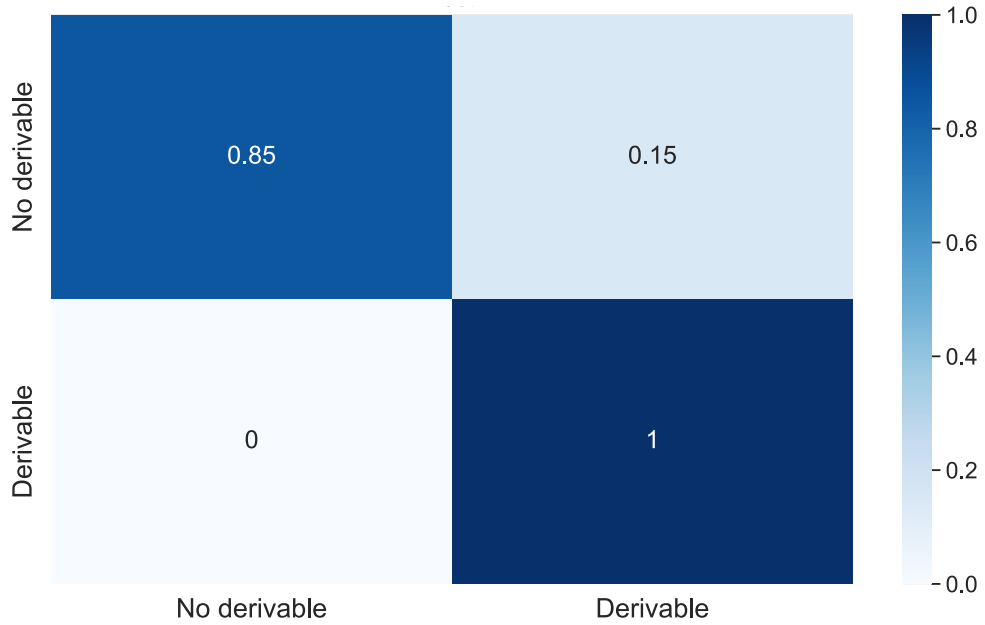


Figura 4.19 Matriz de confusión de los casos derivables de la RD sobre el conjunto de test.

Capítulo 5

Discusión

Los resultados obtenidos permiten proponer un método automático de diagnóstico general de patologías oculares basado en la arquitectura CNN ResNet-50 y las técnicas *data augmentation*, *dropout*, *transfer learning* y *fine-tuning*. El método se ha desarrollado y probado en una gran variedad de imágenes de fondo de ojo disponibles públicamente.

En este capítulo se realiza una valoración crítica de los resultados obtenidos con el método desarrollado. Se van a discutir los resultados obtenidos a partir de la BD empleada este trabajo. Para ello, los resultados se contrastan con los obtenidos en estudios previos y se comprueba la utilidad del estudio.

5.1. Clasificación de la severidad de la Retinopatía Diabética

Para la realización del sistema desarrollado se han empleado distintas arquitecturas con el objetivo de determinar cuál es la óptima. En la fase de entrenamiento se han configurado los parámetros del modelo. Se ha escogido la mejor arquitectura en base a los resultados obtenidos sobre el conjunto de validación. Finalmente, sobre el conjunto de test, que contiene imágenes independientes, se puede comprobar la capacidad de generalización de la red.

En este primer escenario de clasificación de la severidad de la RD se puede observar que los resultados obtenidos en la fase de test son válidos. En cuanto a la precisión del sistema, esta es de un 94.10%. La precisión es el porcentaje de clases clasificadas correctamente respecto al total de datos existentes. En el contexto de clasificación multiclase, resulta más adecuado calcular el coeficiente de kappa ya que tiene en cuenta también el efecto del azar. Además, se trata de una medida más útil para usar en problemas que tienen un desequilibrio entre clases, como ocurre en este caso. Como se mostró anteriormente, el coeficiente kappa obtenido ha sido de 0.919. Esta métrica suele variar de 0, que indica un

acuerdo aleatorio entre calificadores, a 1, que muestra un acuerdo total entre calificadores. En este caso, los calificadores son los expertos oftalmólogos y la predicción del modelo. Por tanto, el valor obtenido indica una alta concordancia entre los valores predichos por el sistema y las anotaciones reales.

Debido a la dificultad de analizar el rendimiento del modelo, se han complementado los resultados con la matriz de confusión obtenida. Esta matriz recoge en la diagonal el porcentaje de imágenes clasificadas correctamente, donde la clase predicha es la misma que la clase definida por un especialista. En la fase de test, la predicción de los pacientes sin RD alcanza un porcentaje de acierto del 98%. Dicho de otro modo, casi todas las retinografías que no contienen patología son detectadas correctamente. Por otra parte, las imágenes que pertenecen a una RDNP moderada son clasificadas con un 87% de acierto, siendo siguiente clase que mayores aciertos presenta. Esto puede justificarse debido al desbalanceo de clases, donde las clases con mayor número de aciertos se corresponden con las que más número de imágenes contienen en la BD. Por otra parte, las clases RDNP leve y RDNP severa son las que menor porcentaje de acierto obtienen en la matriz de confusión, pudiéndose explicar también en términos del número de imágenes de la BD, ya que se trata de las clases con menor número de ellas.

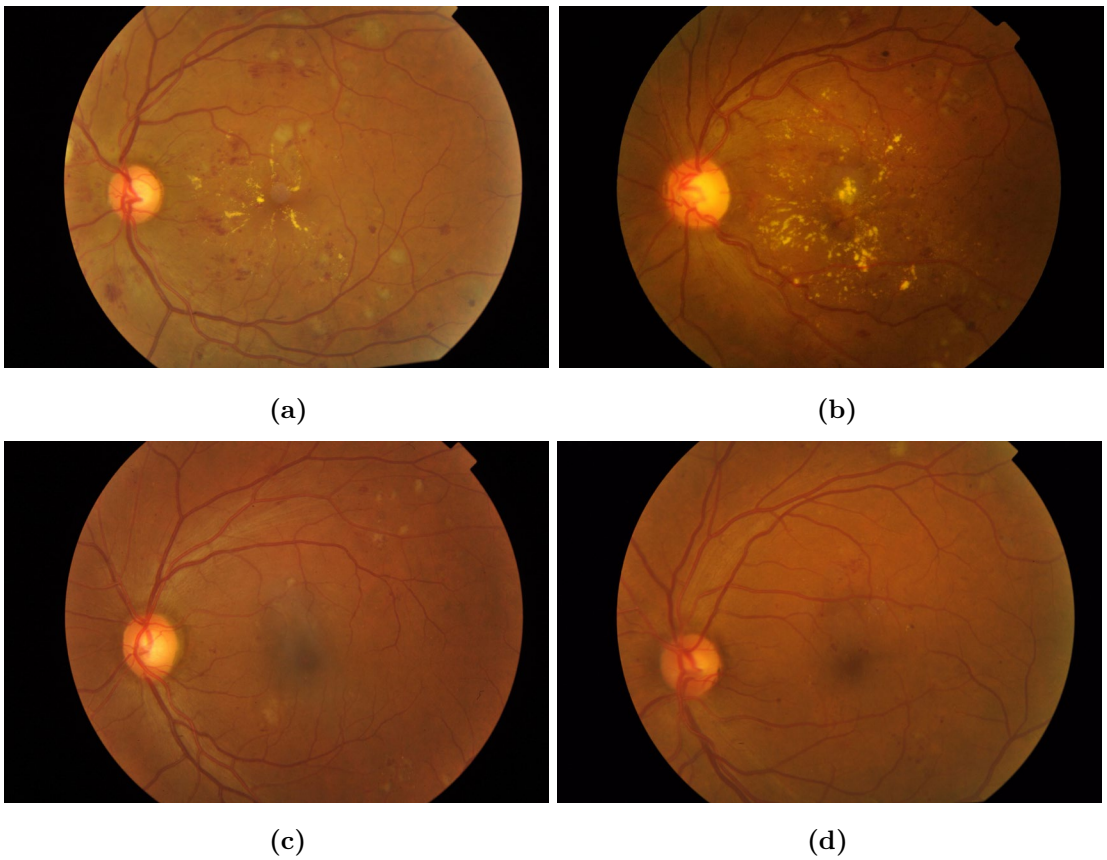
Por otra parte, se han calculado los valores de sensibilidad y especificidad, obteniendo un 66.82% y un 96.31% de media, respectivamente. La sensibilidad se refiere a la capacidad de identificar correctamente las entradas que pertenecen a la clase en cuestión, mientras que la especificidad se refiere a la capacidad de identificar correctamente aquellos casos que no pertenecen a la clase a analizar. Si ahora analizamos los resultados por clases, podemos observar como la sensibilidad en los casos de pacientes sanos alcanza un 97.99%, lo que significa que en un porcentaje muy bajo de los casos el sistema detecta que una imagen pertenece a algún grado de RD cuando en realidad no es así. En este mismo grupo de imágenes se alcanza una especificidad de 98.80%, siendo esta la capacidad del sistema para identificar las imágenes que presentan algún grado de RD. Al analizar estas métricas en los casos de RDNP leve y RDNP severa, se aprecia que los valores de sensibilidad disminuyen considerablemente a un 50% y un 41.18%, respectivamente. Esto significa que el sistema no clasifica correctamente las imágenes pertenecientes a esas clases. El hecho de que las clases que peor sensibilidad obtienen sean estas, puede deberse al gran desbalanceo de la BD, donde las clases de RDNP leve y RDNP severa contienen el 5% y 10% de imágenes de la BD, respectivamente. Uno de los principales problemas que presentan los

conjuntos de datos no balanceados es que la red aprende a distinguir correctamente aquellas clases que contienen más muestras, mientras que las clases minoritarias no se detectan correctamente (Murphey *et ál.*, 2004). Sin embargo, la especificidad obtenida para estas mismas clases (RDNP leve y RDNP severa) es de un 97.02% y 97.13%, lo que indica que el sistema identifica con bastante precisión aquellos casos que no pertenecen a estas clases.

Relacionado con los valores de sensibilidad y especificidad, el último método estadístico empleado para evaluar el sistema desarrollado ha sido la curva ROC. La capacidad discriminativa de un sistema automático de ayuda al diagnóstico se refiere a su habilidad para distinguir pacientes sanos frente a enfermos. En este caso, también resulta necesario discriminar entre los pacientes enfermos el grado de severidad de la RD. Para ello, el parámetro a estimar es el área bajo la curva ROC (AUC), medida única e independiente de la prevalencia de la enfermedad en estudio. El AUC refleja lo bueno que es el modelo para clasificar los distintos grados de severidad. Como se ha mostrado en el capítulo anterior, el AUC de los pacientes sanos sobre el conjunto de test es de un 0.99, lo que indica que un mínimo de imágenes son clasificadas incorrectamente. Asimismo, la segunda clase que alcanza un mayor valor en este parámetro es la perteneciente a la RDNP moderada, con un valor de 0.94. Del mismo modo que se ha justificado en los valores de sensibilidad, estos resultados podrían mostrar una gran dependencia entre el número de imágenes empleadas y la capacidad de generalización de la red.

Las imágenes pertenecientes a la BD APTOS-2019 han sido etiquetadas por diferentes oftalmólogos. En este contexto, la discordancia en la clasificación de las imágenes por parte de distintos especialistas hace que la misma imagen pueda ser enmarcada en dos grados distintos de la RD. En este sentido, el hecho de que la arquitectura haya fallado en la predicción de la clase RDNP severa no es casualidad. Esto se puede deber a que en el entrenamiento la red haya aprendido a detectar como RDNP moderada una imagen muy parecida a la que en el conjunto de test es RDNP severa. En la Figura 5.1, se muestran dos ejemplos de imágenes de RDNP severa junto con las posibles imágenes pertenecientes a RDNP moderada con las que se podrían haber confundido dada su gran similitud. Esta clasificación errónea también se da en las clases RDNP leve y RDNP moderada. Tal y como pasaba con las dos clases anteriores, es posible que una imagen que en el conjunto de entrenamiento se clasifica como RDNP moderada sea muy

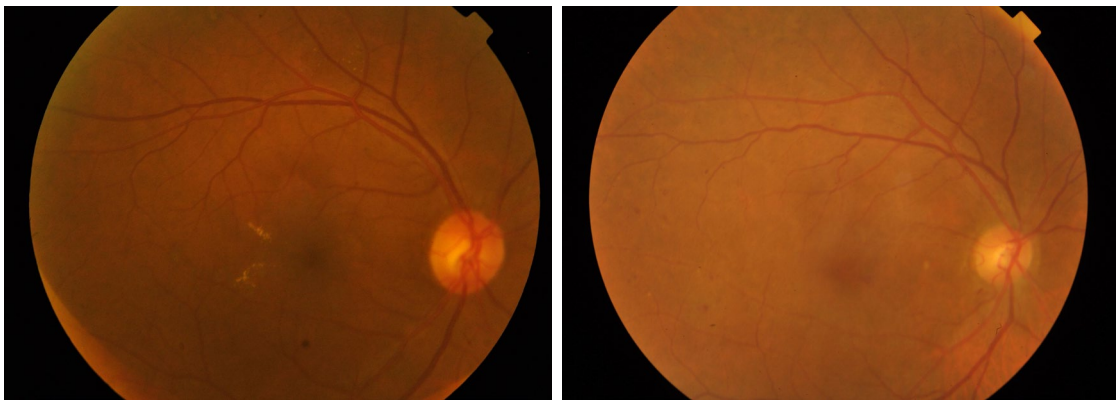
similar a una que en el conjunto de test pertenece a la clase RDNP leve. En la Figura 5.2 se ilustra este caso.



(a) **(b)**

(c) **(d)**

Figura 5.1 **(a)** Imagen perteneciente a la clase RDNP severa del conjunto de test. **(b)** Imagen perteneciente a la clase RDNP moderada del conjunto de entrenamiento. **(c)** Imagen perteneciente a la clase RDNP severa del conjunto de test. **(d)** Imagen perteneciente a la clase RDNP moderada del conjunto de entrenamiento.



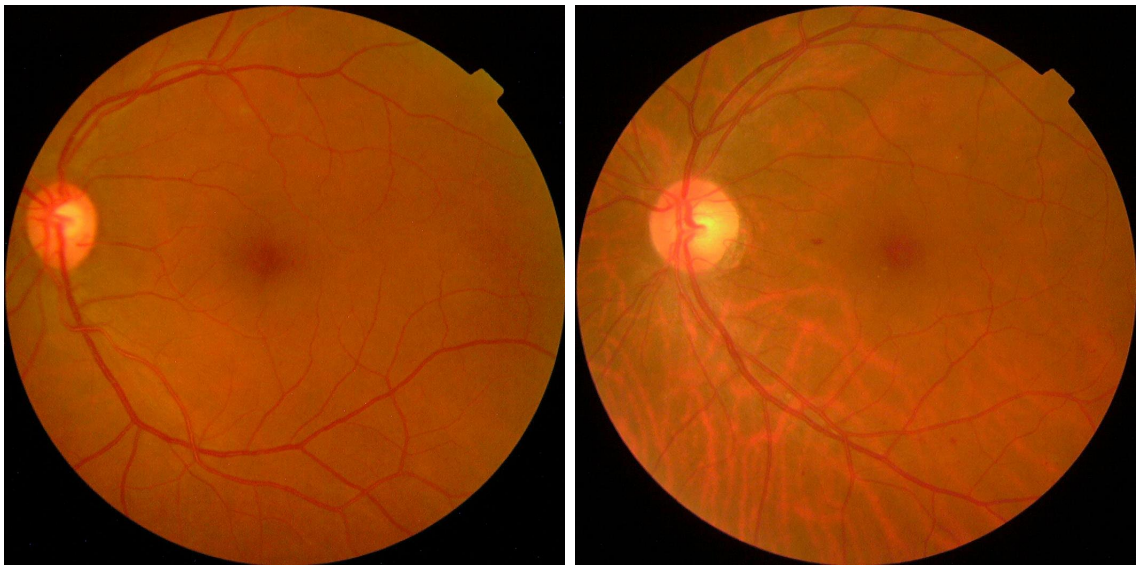
(a) **(b)**

Figura 5.2 **(a)** Imagen perteneciente a la clase RDNP leve del conjunto de test. **(b)** Imagen perteneciente a la clase RDNP moderada del conjunto de entrenamiento.

5.2. Detección de la presencia de la Retinopatía Diabética

En este escenario se analiza la capacidad del sistema para detectar la presencia de RD en retinografías. El análisis de los resultados se realiza sobre el conjunto de test, ya que muestra la capacidad de generalización del sistema a otras imágenes no empleadas durante el entrenamiento. En este conjunto de imágenes se puede observar que el valor de sensibilidad obtenido ha sido de 97.99%. Tal y como está construida la matriz de confusión, este valor indica la capacidad del sistema para detectar las imágenes pertenecientes a pacientes sanos. Por otra parte, el valor de especificidad alcanza su valor máximo (100%) lo que significa que en ningún caso el sistema detecta que una imagen no pertenece a un paciente con RD cuando verdaderamente sí pertenece a esa clase. Dicho de otro modo, todas las retinografías que contienen patología son detectadas correctamente. Por otra parte, el valor de precisión es de 98.91%. Como ya se ha comentado, la precisión indica el número de aciertos con respecto al total de datos. Como no existen errores con la detección de las imágenes patológicas, todos ellos se cometen con las que no contienen signos patológicos. En concreto se han clasificado incorrectamente 4 imágenes del total de 366 que conforman el conjunto de test. Para obtener más información sobre el rendimiento de la clasificación, se ha analizado la curva ROC. Esta curva muestra la relación entre la sensibilidad (tasa de verdaderos positivos) y la especificidad (tasa de falsos positivos). También se ha comprobado que el AUC es un estándar de facto en los problemas de detección en el ámbito médico. En este estudio se ha obtenido un valor de AUC de 1, lo que demuestra que el sistema desarrollado ofrece un gran rendimiento.

Del mismo modo que ocurría en el anterior escenario, el hecho de que el sistema detecte como patológica una imagen que en realidad no lo es puede deberse a que en la BD se pueden encontrar imágenes parecidas etiquetadas por los expertos dentro de distintas clases. Esto puede haber provocado que en el entrenamiento la red haya aprendido a detectar como patológica una imagen muy parecida a la que pertenece a un sujeto sano en el conjunto de test. Por ejemplo, en la Figura 5.3, se muestra una imagen perteneciente a un sujeto sano del conjunto de test junto con una posible imagen patológica perteneciente al conjunto de entrenamiento con la que se podría haber confundido dada su gran similitud.



(a) **(b)**

Figura 5.3 **(a)** Imagen perteneciente a un sujeto sano del conjunto de test. **(b)** Imagen perteneciente a un sujeto con RD del conjunto de entrenamiento.

5.3. Clasificación de los casos derivables y no derivables de la Retinopatía Diabética

El último de los escenarios planteados en este trabajo se corresponde con la división de la BD en casos derivables y casos no derivables de la RD. Se ha considerado que los casos no derivables pertenecen a los casos de ausencia de RD y RDNP leve, mientras que los casos derivables son aquellos que presentan RDNP moderada, RDNP severa y RD proliferativa.

El análisis de los resultados se realiza sobre el conjunto de test, lo que indica la capacidad de generalización del sistema a imágenes nuevas, no empleadas en la fase de entrenamiento. En este conjunto de imágenes se puede observar que el valor de sensibilidad obtenido ha sido de 85.15%. Tal y como está construida la matriz de confusión, este valor indica la capacidad del sistema para detectar las imágenes pertenecientes a los casos no derivables. Por otra parte, el valor de especificidad alcanza el valor de 100%, lo que significa que en ningún caso el sistema detecta que una imagen no pertenece a un caso derivable cuando verdaderamente sí se trata de un caso derivable. Esto significa que todas las retinografías que pertenecen a casos derivables son detectadas correctamente. Por otra parte, el valor de precisión es de 90.71%. Al no existir errores en la detección

de los casos derivables, todos ellos se cometen con en los casos no derivables. En concreto se han clasificado incorrectamente 34 imágenes del total de 366 que conforman el conjunto de test. En este escenario también se ha calculado el valor de AUC, siendo este de 0.97, lo que muestra que el sistema tiene una gran capacidad de detección de los casos derivables y no derivables de la RD.

5.4. Comparativa con estudios previos

Se ha realizado una comparativa de los resultados obtenidos en este TFM con los obtenidos en estudios previos. Esta comparativa se ha dividido en tres tablas, que se corresponden con los tres escenarios planteados en este estudio.

La comparación realizada para la clasificación de la severidad de la RD se puede observar en la Tabla 5.1. La mayoría de los estudios previos miden sus resultados en base al coeficiente kappa, pero también se encuentran estudios que proporcionan los valores de sensibilidad, especificidad, AUC y precisión. En este TFM se han incluido todas estas métricas, y, adicionalmente, las matrices de confusión para poder analizar los resultados para cada una de las clases con mayor detalle.

Entre los estudios analizados se puede observar cómo la mayoría de los que se han incluido utilizan la misma BD. Esto permite realizar la comparación de una manera más precisa debido a que el grupo de imágenes es el mismo. También se puede observar que los estudios que emplean una BD con un número de imágenes mayor, como es el caso de EyePACS o la BD privada utilizada en Galdran *et ál.* (2019), obtienen unos resultados ligeramente mejores a los obtenidos en este TFM.

Como se puede observar, el método propuesto es el que mayor coeficiente kappa, especificidad y precisión presenta. Esto significa que la red es capaz de predecir con mayor precisión el grado de RD respecto a otros métodos previamente publicados. Por otra parte, la mayor parte de los estudios no incluyen valores de AUC, con lo que es difícil realizar comparaciones en base a esta métrica. En este sentido, se puede destacar el estudio de Galdran *et ál.* (2019), donde obtiene una AUC de más del 0.9.

A la vista de los resultados obtenidos, se puede considerar que el método propuesto podría ser de utilidad en un entorno clínico, facilitando la clasificación automática de la RD.

Autor/res, año	**Descripción breve del método**	**Resultados sobre el conjunto de test**				
		Coeficiente kappa	**Sensibilidad**	**Especificidad**	**AUC**	**Precisión**
Krause *et ál.* (2018)	CNN con arquitectura Inception-V3	0.84 (EyePACS-2)	97.1% (EyePACS-2)	92.3% (EyePACS-2)	-	-
Dekhil *et ál.* (2019)	CNN con *transfer learning*	0.78 (APTOS-2019)	-	-	-	77%
Galdran *et ál.* (2019)	CNN con *Label Smoothing*	0.777 (BD privada)	-	-	0.922 (BD privada)	-
Kassani *et ál.* (2019)	Xception con *transfer learning*	-	88.24% (APTOS-2019)	87% (APTOS-2019)	0.918 (APTOS-2019)	83.09% (APTOS -2019)
Bodapati *et ál.* (2020)	CNN	0.711 (APTOS-2019)	-	-	-	81.7% (APTOS-2019)
Pham *et ál.* (2020)	EfficientNet-B5 con *transfer learning*	0.90 (APTOS-2019)	-	-	-	-
Majumder & Kehtarnavaz (2021)	Xception con *transfer learning*	0.90 (APTOS-2019) 0.88 (EyePACS-2)	-	-	-	86% (APTOS-2019) 82% (EyePACS -2)
Método propuesto	Normalización del color, *data augmentation*, *transfer learning*, ResNet-50 y *fine tuning*	**0.919** (APTOS-2019)	66.82% (APTOS-2019)	**96.31%** (APTOS-2019)	**0.93** (APTOS-2019)	**94.10%** (APTOS-2019)

Tabla 5.1 Comparación de los resultados obtenidos en la clasificación de la severidad de la RD con estudios anteriores.

Respecto a la detección de la presencia de RD, los resultados obtenidos permiten proponer un método automático de detección de la presencia de la RD. En este contexto, es posible establecer una comparación con estudios previos, como se muestra en la Tabla 5.2. Cabe destacar que los resultados obtenidos en este TFM superan a los obtenidos en otros estudios en términos de sensibilidad, especificidad, AUC y precisión. No obstante, en este caso, no todos los trabajos han utilizado las mismas bases de datos, por lo que la comparación debe hacerse con precaución. En el trabajo de Choi *et ál.* (2017) se utilizó la arquitectura VGG-19 encadenada a un clasificador *random forest* junto con *transfer learning*. Sin embargo, no se aplicó la técnica *data augmentation*. El estudio de Zhang *et ál.* (2019), realizado con una BD de 13.767 imágenes, también emplea una CNN con *transfer learning*. Además, es posible que los buenos resultados de este estudio también se deban en gran parte al gran número de imágenes empleadas durante el entrenamiento. Finalmente, el trabajo de Skouta *et ál.* (2021) contiene 2400 imágenes, y emplea una CNN basada en la arquitectura VGG, obteniendo unos resultados ligeramente inferiores al método desarrollado. Por otra parte, es necesario destacar los métodos de Koh *et ál.*, que se basaron en técnicas clásicas de procesado de imagen, lo que permite confirmar que las técnicas de *deep learning* pueden superar los resultados obtenidos por los métodos convencionales.

Por último, la comparación realizada para la clasificación de los casos derivables y no derivables se puede observar en la Tabla 5.3. En este caso, tampoco todos los trabajos han utilizado las mismas bases de datos, por lo que la comparación debe hacerse con cautela. Es posible observar que el estudio de Chetoui & Akhloufi (2020) realizado sobre la BD EyePACS obtiene la sensibilidad más alta de todos los trabajos expuestos. Esta BD contiene más de 35000 imágenes, frente a las 3662 imágenes empleadas en este TFM. Los resultados sobre BBDD que contienen un número inferior de imágenes son inferiores a los obtenidos en este trabajo. Este es el caso expuesto en Saxena *et ál.* (2020), donde se utiliza la BD Messidor-2 con únicamente 1748 imágenes, obteniendo una sensibilidad de 81.02%. Respecto a los valores de especificidad, el sistema propuesto proporciona los resultados más altos (100%) lo que significa que detecta todos los casos de RD derivable. Si se comparan los valores de AUC, se puede comprobar que todos los trabajos obtienen resultados similares, siendo el trabajo de Gulshan *et ál.* (2016) el que mayor rendimiento presenta en términos de dicha métrica. Por último, no ha sido posible realizar una comparación en términos de precisión ni coeficiente

kappa debido a que ninguno de los estudios ha incluido los valores obtenido sobre estas métricas.

Autor/res, año	**Descripción breve del método**	**Resultados sobre el conjunto de test**			
		Sensibilidad	**Especificidad**	**AUC**	**Precisión**
Choi *et ál.* (2017)	VGG-19, *transfer learning*, clasificador *random forest*	80.30% (STARE)	85.80% (STARE)	0.903 (STARE)	-
Koh *et ál.* (2017)	PHOG, SURF, AdaSyn, clasificador k-NN	95.00% (BD privada)	97.42% (BD privada)	-	96.21% (BD privada)
Koh *et ál.* (2018)	PHOW, vector de Fisher, clasificador *random forest*	96.73% (BD privada)	96.96% (BD privada)	-	96.79% (BD privada)
Zhang *et ál.* (2019)	CNN con *transfer learning* y *ensemble learning*	97.5% (BD privada)	97.7% (BD privada)	0.977 (BD privada)	97.67% (BD privada)
Skouta *et ál.* (2021)	CNN basada en VGG	96.5% (BD basada en EyePACs)	94.5% (BD basada en EyePACs)	-	95.5% (BD basada en EyePACs)
Método propuesto	Normalización del color, *data augmentation*, *transfer learning*, ResNet-50 y *fine tuning*	**97.99%** (APTOS-2019)	**100%** (APTOS-2019)	**1** (APTOS -2019)	**98.91%** (APTOS-2019)

Tabla 5.2 Comparación de los resultados obtenidos en la detección de la RD con estudios anteriores.

Es necesario destacar que la mayoría de los trabajos presentados tanto en la detección de la presencia de la RD como en la división en casos derivables y no derivables se han realizado a través de sistemas de clasificación binaria. En este TFM, se ha empleado el modelo multiclase y se han agrupado las clases para obtener los resultados de estos grupos. Es posible que el entrenamiento del modelo adaptado a una clasificación binaria hubiese proporcionado unos resultados mejores en estos dos escenarios.

Autor/res, año	**Descripción breve del método**	**Resultados sobre el conjunto de test**		
		Sensibilidad	**Especificidad**	**AUC**
Gulshan *et ál.* (2016)	CNN	90.3% (EyePACS-1)	98.1% (EyePACS-1)	0.991 (EyePACS-1)
Ting *et ál.* (2017)	CNN	90.5% (BD privada)	91.6% (BD privada)	0.936 (BD privada)
Sahlsten *et ál.* (2019)	Inception-V3 con *transfer learning*	89.6% (BD privada)	97.4% (BD privada)	0.987 (BD privada)
Chetoui & Akhloufi (2020)	EfficientNET	91.7% (EyePACS) 91.4% (APTOS-2019)	98.9% (EyePACS) 97.2% (APTOS-2019)	0.984 (EyePACS) 0.966 (APTOS-2019)
Saxena *et ál.* (2020)	CNN	88.84% (Messidor–1) 81.02% (Messidor–2)	89.92% (Messidor–1) 86.09% (Messidor–2)	0.958 (Messidor–1) 0.92 (Messidor–2)
Método propuesto	Normalización del color, *data augmentation*, *transfer learning*, ResNet-50 y *fine tuning*	85.15% (APTOS-2019)	**100%** (APTOS-2019)	0.97 (APTOS-2019)

Tabla 5.3 Comparación de los resultados obtenidos en la clasificación de los casos derivables y no derivables de la RD con estudios anteriores.

5.5. Interpretación mediante SHAP

En las aplicaciones médicas, es importante poder interpretar las predicciones de los modelos. Aunque unos buenos resultados sobre los conjuntos de validación y test pueden ser una medida para seleccionar un modelo óptimo, resulta insuficiente para el uso de este modelo en las aplicaciones reales. En un entorno clínico real se necesita entender en qué se basa el modelo para tomar las decisiones.

Mediante el uso de SHAP es posible visualizar características que contribuyen a la evaluación de cada etapa de la enfermedad. El uso de SHAP permite asegurar que el modelo aprende características útiles durante el entrenamiento, así como que utiliza características correctas en el momento de la inferencia. Además, en casos inciertos, la visualización de las características más destacadas puede ayudar al médico a centrarse en las regiones de interés en las que las características son más notables.

En la Figura 5.4, se muestran varios ejemplos de visualización de los valores SHAP para distintas imágenes del conjunto de test. Esta figura representa, de izquierda a derecha, la imagen original y los mapas SHAP para cada clase, ordenados por orden de severidad, siendo la imagen de la izquierda la que representa la ausencia de enfermedad, y la imagen de la derecha la que representa la RD proliferativa.

La interpretación proporcionada por la visualización basada en SHAP es la siguiente: el color rojo indica los píxeles que aumentan el valor de salida para una clase determinada, y el color azul indica las características que disminuyen el valor de salida para una clase determinada. Esto significa que los pixeles en rojo representan píxeles que pertenecen a una clase concreta, mientras que los pixeles azules lo descartan. Por ejemplo, en la Figura 5.4(a) se puede apreciar como la clase que indica la ausencia de RD es la que mayor porcentaje de pixeles rojos presenta. Estos pixeles se encuentran en una región que no contiene lesiones aparentes, y que podría ser, en determinadas condiciones, también confundida con RDNP leve. La Figura 5.4(b) muestra un caso de RDNP leve. En este caso, se puede observar como aparecen tres regiones coloreadas, pudiendo indicar la presencia de algún cambio de textura en la imagen, y, por consiguiente, la presencia de alguna lesión. Estas mismas zonas se muestran también en los mapas SHAP de los casos de ausencia de RD y RDNP moderada, clases con las que habitualmente se puede confundir la RDNP leve. En el caso de la Figura 5.4(d)

se aprecia que a medida que aumenta el grado de severidad de la RD, los pixeles rojos aumentan. Sin embargo, el modelo no asigna ningún pixel rojo a la clase RD proliferativa e indica que esta es la clase a la que pertenece la imagen. Esto puede deberse a que, en el resto de los casos, el sistema identifica zonas que no pertenecen a esa clase (contribuciones SHAP negativas en los pixeles azules), mientras que esta clase no contiene dichos pixeles azules, y por tanto, el sistema no tiene incertidumbre en la predicción.

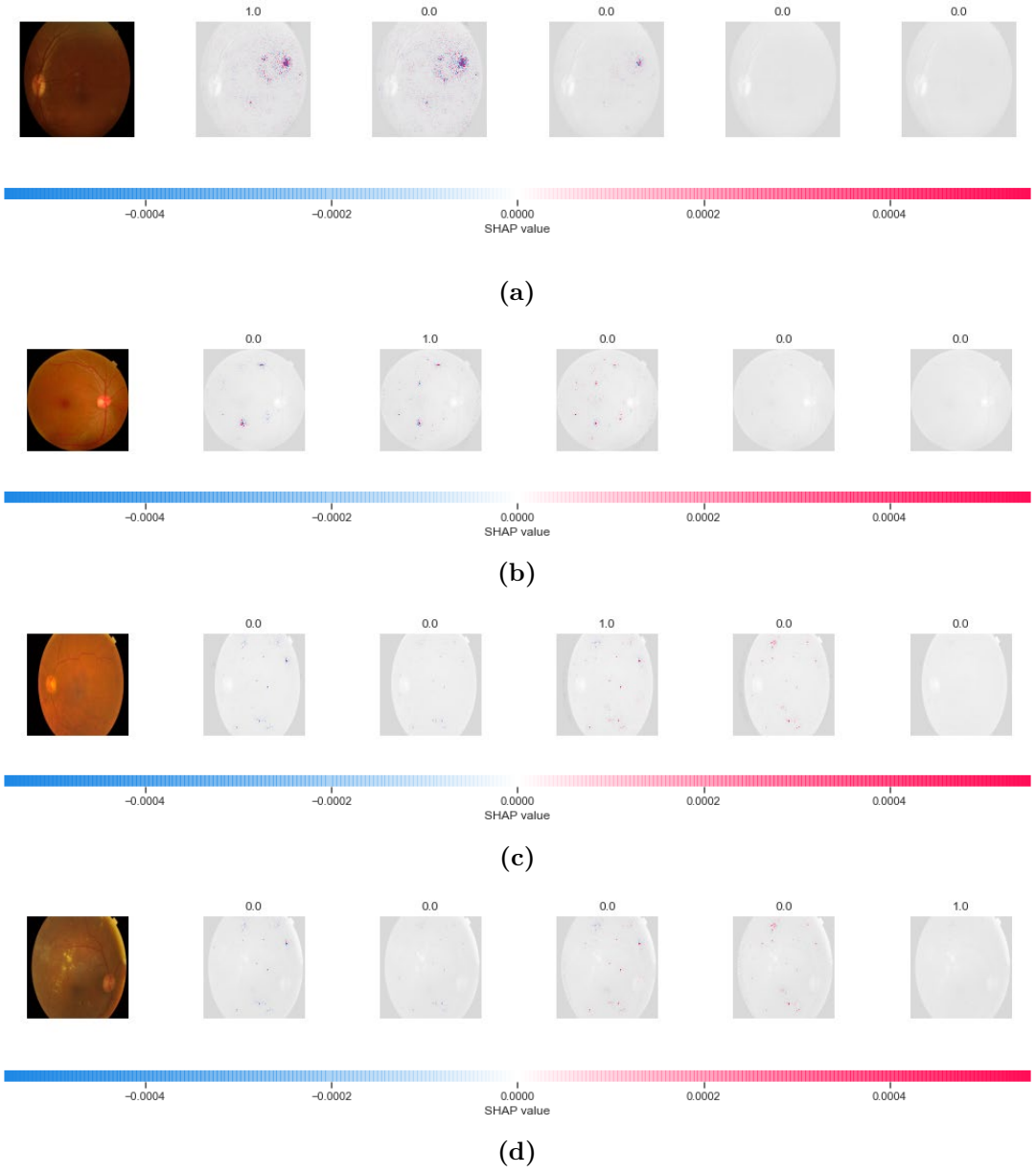


Figura 5.4 Ejemplos de valores SHAP sobre imágenes de fondo de ojo. **(a)** Sin patología. **(b)** RDNP leve. **(c)** RDNP moderada. **(d)** RD proliferativa.

Capítulo 6

Conclusiones y líneas futuras

En plena era de los datos y la automatización, la medicina no puede quedar atrás. El método desarrollado es solo un ejemplo de cómo el *deep learning* puede ayudar a los especialistas a detectar posibles enfermedades en estadios muy tempranos, lo que permitirá tratarlas antes de que puedan afectar a la vida diaria de los pacientes.

En este trabajo se ha comprobado cómo la RD, una de las principales causas de ceguera en todo el mundo, podría detectarse de forma muy temprana, pudiendo así tratarse antes de que avance a sus estadios más graves.

En este sentido, debido a la creciente incidencia de la diabetes, el número de imágenes que tienen que analizar los oftalmólogos es enorme. Además, se trata de una tarea que requiere mucho tiempo y esfuerzo, provocando que el coste de los exámenes oculares y la escasez de especialistas hagan que muchos pacientes no reciban un tratamiento adecuado. Es en este contexto donde el desarrollo de sistemas automáticos que permitan clasificar el grado de severidad de la RD ofrece muchos beneficios. El método desarrollado podría ser empleado en un programa de cribado o *screening*, permitiendo el examen de un mayor número de imágenes en menor tiempo. Del mismo modo, se podría reducir significativamente la carga de trabajo de los especialistas, debido a que permitiría seleccionar únicamente aquellas imágenes que presenten ciertos signos clínicos que requieran un examen adicional.

En este capítulo se resumen las principales contribuciones y conclusiones de este TFM. En la sección 6.1 se han recogido las aportaciones originales de este trabajo. A continuación, en la sección 6.2 se presentan las principales conclusiones extraídas de la aplicación del método desarrollado a la clasificación de la RD. Finalmente, la sección 6.3 se dedica a presentar las limitaciones y posibles líneas de trabajo futuro.

6.1. Contribuciones originales

En este TFM se ha presentado un sistema de análisis automatizado de fotografías de fondo de ojo, aplicadas al diagnóstico asistido por ordenador de la RD. Esta enfermedad constituye una de las principales causas de ceguera prevenible a nivel mundial, y su detección utilizando imágenes de fondo de ojo es valiosa teniendo en cuenta el bajo costo y la naturaleza no invasiva de esta modalidad de imagen.

En el Capítulo 3 se ha descrito en detalle el método desarrollado. Para la implementación se han comparado cuatro arquitecturas diferentes, con el objetivo de tratar de mejorar los resultados de otros métodos centrados en la clasificación de la RD previamente publicados. Entre las contribuciones más importantes del trabajo realizado se pueden destacar:

1. Aplicación de técnicas de preprocesado de imágenes con el objetivo de adaptar la imagen de entrada al sistema de clasificación basado en *deep learning*. El preprocesado de las imágenes se ha basado en una normalización del color y una reducción de la resolución para acelerar el proceso de entrenamiento. La aplicación de estas técnicas permite utilizar el mismo modelo de clasificación sobre datos pertenecientes a distintas BBDD.

2. Desarrollo de un método automático de clasificación de las etapas de la RD a partir del análisis de imágenes de fondo de ojo. Gracias a este sistema, se consigue clasificar retinografías procedentes de diferentes pacientes diabéticos en cinco categorías: sin patología, RDNP leve, RDNP moderada, RDNP severa y RD proliferativa. Para la implementación del sistema se ha hecho uso de CNNs complementadas con otras técnicas como *data augmentation, dropout, transfer-learning* y *fine-tuning*.

3. Evaluación de los resultados obtenidos empleando diversas métricas: coeficiente kappa, sensibilidad, especificidad, AUC, precisión y matrices de confusión. En la literatura, la mayoría de los estudios no incluyen todos estos parámetros. Este trabajo presenta una evaluación exhaustiva sobre el sistema desarrollado.

4. Interpretación de los resultados obtenidos mediante una visualización basada en SHAP.

6.2. Conclusiones

En este trabajo se ha abordado una tarea importante en el ámbito del procesado automático de retinografías para el diagnóstico de la RD: la clasificación de las mismas en sus respectivas etapas del avance de la enfermedad. Los resultados obtenidos indican que **el método propuesto es adecuado para la clasificación de la severidad de la RD y puede suponer una ayuda significativa para los oftalmólogos en la clasificación de la enfermedad y en la práctica clínica, reduciendo los costes y el tiempo para obtener un diagnóstico**.

Junto con esta conclusión principal, se pueden extraer las siguientes conclusiones generales del trabajo presentado:

1. Se ha llevado a cabo un estudio exhaustivo de los diferentes métodos de clasificación que se han empleado en el análisis de imágenes médicas y, en concreto, en la clasificación de la RD. En este sentido, los métodos de procesado de retinografías basados en CNNs son especialmente relevantes para el diseño de sistemas de clasificación automática del grado de severidad de la RD.

2. Se ha llevado a cabo la familiarización con la BD seleccionada. La BD APTOS-2019 consta de 3662 imágenes que han sido clasificadas por distintos oftalmólogos especialistas en cinco clases distintas, que indican el grado de severidad de la RD. Asimismo, las imágenes incluyen diferentes características, de modo que se muestra la gran variabilidad de las imágenes que pueden aparecer en un entorno clínico. Esta BD podrá ser empleada en futuros estudios y permitirá realizar comparaciones objetivas de los resultados obtenidos con diferentes técnicas de clasificación de la RD en retinografías.

3. El funcionamiento de los métodos propuestos se evaluó sobre un conjunto de imágenes independiente y con características variables. La forma de evaluación empleada en este trabajo permite una visión más global acerca del funcionamiento de los métodos debido a que se muestran distintos parámetros. En la mayoría de los trabajos previos, los métodos desarrollados se han evaluado empleando únicamente el coeficiente kappa. En este sentido, el empleo de un único parámetro no permite evaluar con fiabilidad el desempeño del sistema.

4. En los métodos que utilizan *deep learning* resulta imposible determinar cómo llegan las redes neuronales a una conclusión. En este contexto, la visualización con SHAP permite comprender mejor cómo el sistema realiza las predicciones, mostrando qué píxeles de la imagen se identifican con cada grado de severidad de la RD.

5. Los métodos desarrollados son completamente automáticos una vez que se han encontrado los parámetros óptimos a utilizar en cada una de las etapas. Esto significa que no es necesaria la intervención del usuario en ningún momento del análisis de la imagen.

6. Las imágenes digitales brindan ventajas en cuanto al almacenamiento e intercambio entre distintas organizaciones. De esta manera, un sistema automático como el propuesto podría emplearse en cualquier entorno clínico, ofreciendo un seguimiento más rápido de la RD en dos contextos distintos.

    - En un programa de *screening*, sería capaz de analizar un elevado número de imágenes de manera automática y sin la necesidad de la supervisión de un especialista.

    - En el ámbito clínico, sería una herramienta de ayuda al diagnóstico que reduciría la carga de trabajo de los especialistas proporcionando una primera clasificación de la imagen.

    El método expuesto, por tanto, permitiría acortar el tiempo de obtención de un diagnóstico, reducir la carga de trabajo de los oftalmólogos y, como consecuencia, los costes económicos asociados al tratamiento de la RD.

6.3. Limitaciones y líneas futuras

A pesar de que los resultados obtenidos son satisfactorios y similares a los de estudios previos, este estudio no está exento de limitaciones. Algunas de ellas se exponen a continuación:

1. La principal limitación del presente trabajo es el tamaño de la BD con la que se ha trabajado. Para poder extender las conclusiones del estudio realizado sería necesario completar la BD con un mayor número de imágenes. Además, sería necesario que estas nuevas retinografías que formasen parte de la BD fueran más representativas de la variabilidad

existente en la práctica clínica, es decir, utilizando imágenes de distintos retinógrafos o lugares de captura y tomadas por distintos especialistas. Esta heterogeneidad dotaría de un carácter más universal al método y permitiría comprobar la efectividad del mismo con mayor precisión. Cabe esperar que, si se aumenta el número de imágenes, los resultados mejorarían ya que las redes CNN generalizan mejor cuando se entrenan con más datos.

2. Otra de las principales limitaciones del trabajo es el gran desbalanceo existente entre las clases de la BD empleada. Mientras que la clase predominante, las imágenes de retinas sanas, contenía el 50% del total de imágenes; la clase minoritaria, las imágenes de retinas con RDNP severa, contenía el 5% del total. De manera general, las redes neuronales tienden a favorecer las clases mayoritarias, provocando una menor precisión en la clasificación de las clases minoritarias. Por este motivo, la clasificación de la clase RDNP severa ha sido la que ha obtenido las métricas más bajas.

3. También encontramos limitaciones relacionadas con el uso de redes neuronales profundas ya que el coste computacional de estos modelos es enorme debido a la gran cantidad de iteraciones y de datos que es necesario almacenar.

Teniendo en cuenta estas limitaciones, se proponen algunas líneas futuras de investigación tanto centradas en realizar ciertas modificaciones sobre el método propuesto como en otros trabajos de investigación de la misma rama. En este contexto, se plantean:

1. Aumentar el número de imágenes de la BD con el objetivo de evaluar la utilidad diagnóstica del método desarrollado sobre imágenes que presenten más variabilidad en cuanto a color, contraste y calidad, con diferentes resoluciones y grados de apertura del FOV. Por ello, se plantea explorar las posibilidades de adaptar el método desarrollado a una mayor variabilidad de imágenes ampliando la BD disponible durante el entrenamiento. Asimismo, sería deseable que esta nueva BD se encuentre balanceada, de manera que se puedan eliminar los errores asociados al desbalanceo de datos.

2. Realizar ciertas modificaciones sobre el método propuesto para intentar mejorar los resultados obtenidos hasta este momento. Estos resultados

indican que aún hay posibilidad de mejorar el valor de sensibilidad obtenido sobre determinadas clases. Una forma de realizar esto es empleando métodos más complejos en la etapa de preprocesado de imágenes, como el aumento del contraste de las imágenes, lo que permitiría que las lesiones se diferenciasen del fondo de una manera más clara. Asimismo, el valor empleado para algunos hiperparámetros puede perfeccionarse mediante una búsqueda más exhaustiva.

3. Implementar una aplicación que permita la evaluación automática de la severidad de la RD. Esta aplicación sería de gran utilidad en escenarios clínicos y de cribado.

4. Desarrollar un sistema que permita detectar las lesiones de cada etapa de la RD. La severidad de la RD está directamente relacionada con el número de lesiones presentes y con la localización de las mismas. En este sentido, se plantea el desarrollo de un sistema que combine ambos objetivos.

Los resultados obtenidos sugieren que el método desarrollado puede resultar útil para la ayuda al diagnóstico de la RD como un sistema de cribado que permitiría seleccionar únicamente aquellas imágenes que requieran un examen adicional por parte del oftalmólogo, reduciendo de forma significativa su carga de trabajo. Asimismo, permitiría el examen de un mayor número de imágenes en menor tiempo.

Referencias

Abramoff, M. D., & Suttorp-Schulten, M. S. A. (2005). Web-based screening for diabetic retinopathy in a primary care population: the EyeCheck project. *Telemedicine Journal and E-Health : The Official Journal of the American Telemedicine Association*, *11*(6), 668—674. https://doi.org/10.1089/tmj.2005.11.668

Abràmoff, M D, & Niemeijer, M. (2015). Mass Screening of Diabetic Retinopathy Using Automated Methods. *Teleophthalmology in Preventive Medicine*, 41–50. https://doi.org/10.1007/978-3-662-44975-2_4

Abràmoff, Michael D;, Lavin, P. T., Birch, M., Shah, N., & Folk, J. C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *Npj Digital Medicine*, *1*(1), 39. https://doi.org/10.1038/s41746-018-0040-6

Abràmoff, Michael 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 U, R., Chua, C. K., Ng, E. Y. K., Yu, W., & Chee, C. (2008). Application of higher order spectra for the identification of diabetes retinopathy stages. *Journal of Medical Systems*, *32*(6), 481–488. https://doi.org/10.1007/s10916-008-9154-8

Acharya, U R, Lim, C. M., Ng, E. Y. K., Chee, C., & Tamura, T. (2009). Computer-based detection of diabetes retinopathy stages using digital fundus images. *Proceedings of the Institution of Mechanical Engineers. Part H, Journal of Engineering in Medicine*, *223*(5), 545–553. https://doi.org/10.1243/09544119JEIM486

Acharya, U Rajendra, Ng, E. Y. K., Tan, J.-H., Sree, S. V., & Ng, K.-H. (2012). An integrated index for the identification of diabetic retinopathy stages using texture parameters. *Journal of Medical Systems*, *36*(3), 2011–2020. https://doi.org/10.1007/s10916-011-9663-8

Achille, A., & Soatto, S. (2018). Information Dropout: Learning Optimal Representations Through Noisy Computation. *IEEE Transactions on*

Pattern Analysis and Machine Intelligence, *40*(12), 2897–2905. https://doi.org/10.1109/TPAMI.2017.2784440

Adriman, R., Muchtar, K., & Maulina, N. (2021). Performance Evaluation of Binary Classification of Diabetic Retinopathy through Deep Learning Techniques using Texture Feature. *Procedia Computer Science*, *179*, 88–94. https://doi.org/https://doi.org/10.1016/j.procs.2020.12.012

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

Aliseda, D., & Berástegui, L. (2008). Retinopatía diabética. *Anales Del Sistema Sanitario de Navarra*, *31*, 23–34. http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1137-66272008000600003&nrm=iso

Alyoubi, W. L., Abulkhair, M. F., & Shalash, W. M. (2021). Diabetic Retinopathy Fundus Image Classification and Lesions Localization System Using Deep Learning. *Sensors*, *21*(11). https://doi.org/10.3390/s21113704

Asia Pacific Tele-Ophthalmology Society. (2021). *Asia Pacific Tele-Ophthalmology Society – APTOS*. https://asiateleophth.org/

Bar, Y., Diamant, I., Wolf, L., & Greenspan, H. (2015). Deep learning with non-medical training used for chest pathology identification. *Medical Imaging*.

Barry, C. J., McAllister, I. L., Constable, I. J., & Yogesan, K. (2006). Diabetic Retinopathy Screening: What Model is Appropriate? In K. Yogesan, S. Kumar, L. Goldschmidt, & J. Cuadros (Eds.), *Teleophthalmology* (pp. 87–98). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-33714-8_12

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/https://doi.org/10.1016/j.csbj.2016.10.001

Bhardwaj, C., Jain, S., & Sood, M. (2021). Diabetic retinopathy severity grading employing quadrant-based Inception-V3 convolution neural network architecture. *International Journal of Imaging Systems and Technology*, *31*(2), 592–608. https://doi.org/https://doi.org/10.1002/ima.22510

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, Inc.

Bodapati, J. D., Naralasetti, V., Shareef, S. N., Hakak, S., Bilal, M., Maddikunta,

P. K. R., & Jo, O. (2020). Blended Multi-Modal Deep ConvNet Features for Diabetic Retinopathy Severity Prediction. *Electronics*, *9*(6). https://doi.org/10.3390/electronics9060914

Briggs, W. M., Zaretzki, R., Pepe, M. S., & Hand, D. J. (2008). The Skill Plot: A Graphical Technique for Evaluating Continuous Diagnostic Tests. *Biometrics*, *64*(1), 250–261. http://www.jstor.org/stable/25502043

Bronzino, J. D. (2006). *The Biomedical Engineering Handbook*. CRC/Taylor & Francis. https://books.google.com.ni/books?id=NcJzjwEACAAJ

Chetoui, M., & Akhloufi, M. A. (2020). Explainable Diabetic Retinopathy using EfficientNET*. *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 1966–1969. https://doi.org/10.1109/EMBC44109.2020.9175664

Cheung, N., Donaghue, K. C., Liew, G., Rogers, S. L., Wang, J. J., Lim, S.-W., Jenkins, A. J., Hsu, W., Lee, M. L., & Wong, T. Y. (2009). Quantitative Assessment of Early Diabetic Retinopathy Using Fractal Analysis. *Diabetes Care*, *32*(1), 106–110. https://doi.org/10.2337/dc08-1233

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), 1–16. https://doi.org/10.1371/journal.pone.0187336

Chung, Y.-A., & Weng, W.-H. (2017). *Learning Deep Representations of Medical Images using Siamese CNNs with Application to Content-Based Image Retrieval*.

Ciompi, F., de Hoop, B., van Riel, S. J., Chung, K., Scholten, E. T., Oudkerk, M., de Jong, P. A., Prokop, M., & van Ginneken, B. (2015). Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box. *Medical Image Analysis*, *26*(1), 195–202. https://doi.org/10.1016/j.media.2015.08.001

Colas, E., Besse, A., Orgogozo, A., Schmauch, B., Meric, N., & Besse, E. (2016). Deep learning approach for diabetic retinopathy screening. *Acta Ophthalmologica*, *94*. https://doi.org/10.1111/j.1755-3768.2016.0635

Costa, P., & Campilho, A. (2017). Convolutional bag of words for diabetic retinopathy detection from eye fundus images. *IPSJ Transactions on Computer Vision and Applications*, *9*. https://doi.org/10.1186/s41074-017-

0023-6

Crosby-Nwaobi, R., Heng, L. Z., & Sivaprasad, S. (2012). Retinal vascular calibre, geometry and progression of diabetic retinopathy in type 2 diabetes mellitus. *Ophthalmologica. Journal International d'ophtalmologie. International Journal of Ophthalmology. Zeitschrift Fur Augenheilkunde*, *228*(2), 84–92. https://doi.org/10.1159/000337252

Dekhil, O., Naglah, A., Shaban, M., Ghazal, M., Taher, F., & Elbaz, A. (2019). Deep Learning Based Method for Computer Aided Diagnosis of Diabetic Retinopathy. *2019 IEEE International Conference on Imaging Systems and Techniques (IST)*, 1–4. https://doi.org/10.1109/IST48021.2019.9010333

Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 210–215. https://doi.org/10.23919/MIPRO.2018.8400040

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 Suppl), 786–806.

Ege, B. M., Hejlesen, O. K., Larsen, O. V, Møller, K., Jennings, B., Kerr, D., & Cavan, D. A. (2000). Screening for diabetic retinopathy using computer based image analysis and statistical classification. *Computer Methods and Programs in Biomedicine*, *62*(3), 165—175. https://doi.org/10.1016/s0169-2607(00)00065-1

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*, 4. https://doi.org/10.3389/frai.2020.00004

Enderle, J., & Bronzino, J. D. (2011). *Introduction to Biomedical Engineering*. https://doi.org/10.1016/B978-0-12-238662-6.X5000-9

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/https://doi.org/10.1016/j.patrec.2005.10.010

Federación Internacional de Diabetes. (2019). *Atlas de la Diabetes de la FID* (9th ed.). https://www.diabetesatlas.org

Feng, J., He, X., Teng, Q., Ren, C., Chen, H., & Li, Y. (2019). Reconstruction of

porous media from extremely limited information using conditional generative adversarial networks. *Physical Review E*, *100*. https://doi.org/10.1103/PhysRevE.100.033308

Fronzetti, N. (2019). *Predictive Neural Network Applications for Insurance Processes*.

Galdran, A., Chakor, H., Alrushood, A., Kobbi, R., Christodoulidis, A., Chelbi, J., Racine, M.-A., & Benayed, I. (2019). *Automatic classification and triage of diabetic retinopathy from retinal images based on a convolutional neural networks (CNN) method*. *97*. https://doi.org/10.1111/j.1755-3768.2019.5391

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.

Gargeya, R., & Leng, T. (2017). Automated Identification of Diabetic Retinopathy Using Deep Learning. *Ophthalmology*, *124*(7), 962–969. https://doi.org/10.1016/j.ophtha.2017.02.008

Gowda, T., You, W., Lignos, C., & May, J. (2021). *Macro-Average: Rare Types Are Important Too*. https://doi.org/10.18653/v1/2021.naacl-main.90

Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for Multi-Class Classification: an Overview*. http://arxiv.org/abs/2008.05756

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 Recogn.*, *77*(C), 354–377. https://doi.org/10.1016/j.patcog.2017.10.013

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, *316*(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216

Hadjikhani, N., & Tootell, R. B. H. (2000). Projection of rods and cones within human visual cortex. *Human Brain Mapping*, *9*(1), 55–63. https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0193%282000%299%3A1%3C55%3A%3AAID-HBM6%3E3.0.CO%3B2-U

Hattiya, T., Dittakan, K., & Musikasuwan, S. (2021). Diabetic Retinopathy Detection Using Convolutional Neural Network: A Comparative Study on

Different Architectures. *Mahasarakham International Journal of Engineering Technology*, *7*, 11. https://doi.org/10.14456/mijet.2021.8

Haykin, S. (1998). *Neural Networks: A Comprehensive Foundation* (2nd ed.). Prentice Hall PTR.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition*. https://arxiv.org/abs/1512.03385

Ho, Y., & Wookey, S. (2019). The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling. *IEEE Access*, *PP*, 1. https://doi.org/10.1109/ACCESS.2019.2962617

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications*. https://arxiv.org/abs/1704.04861

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2016). *Densely Connected Convolutional Networks*. http://arxiv.org/abs/1608.06993

Jelinek, H. J., Cree, M. J., Worsley, D., Luckie, A., & Nixon, P. (2006). An automated microaneurysm detector as a tool for identification of diabetic retinopathy in rural optometric practice. *Clinical & Experimental Optometry*, *89*(5), 299–305. https://doi.org/10.1111/j.1444-0938.2006.00071.x

Kandel, I., & Castelli, M. (2020). Transfer Learning with Convolutional Neural Networks for Diabetic Retinopathy Image Classification. A Review. *Applied Sciences*, *10*(6). https://doi.org/10.3390/app10062021

Kassani, S. H., Kassani, P. H., Khazaeinezhad, R., Wesolowski, M. J., Schneider, K. A., & Deters, R. (2019). Diabetic Retinopathy Classification Using a Modified Xception Architecture. *2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 1–6. https://doi.org/10.1109/ISSPIT47144.2019.9001846

Kels, B. D., Grzybowski, A., & Grant-Kels, J. M. (2015). Human ocular anatomy. *Clinics in Dermatology*, *33*(2), 140–146. https://doi.org/10.1016/j.clindermatol.2014.10.006

Khan, A., Sohail, A., Zahoora, U., & Saeed, A. (2020). A Survey of the Recent Architectures of Deep Convolutional Neural Networks. *Artificial Intelligence Review*, *53*. https://doi.org/10.1007/s10462-020-09825-6

Klein, R., Knudtson, M. D., Lee, K. E., Gangnon, R., & Klein, B. E. K. (2009).

The Wisconsin Epidemiologic Study of Diabetic Retinopathy XXIII: the twenty-five-year incidence of macular edema in persons with type 1 diabetes. *Ophthalmology*, *116*(3), 497–503. https://doi.org/10.1016/j.ophtha.2008.10.016

Koh, Joel E W, Ng, E. Y. K., Bhandary, S. V, Laude, A., & Acharya, U. R. (2017). Automated detection of retinal health using PHOG and SURF features extracted from fundus images. *Applied Intelligence*, *48*(5), 1379–1393. https://doi.org/10.1007/s10489-017-1048-3

Koh, Joel En Wei, Ng, E., Bhandary, S., Hagiwara, Y., Laude, A., & Acharya, U. R. (2018). Automated retinal health diagnosis using pyramid histogram of visual words and Fisher vector techniques. *Computers in Biology and Medicine*, *92*. https://doi.org/10.1016/j.compbiomed.2017.11.019

Krause, J., Gulshan, V., Rahimy, E., Karth, P., Widner, K., Corrado, G. S., Peng, L., & Webster, D. R. (2018). 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

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*, 84–90.

Le, L., Zheng, Y., Carneiro, G., & Yang, L. (2017). *Deep Learning and Convolutional Neural Networks for Medical Image Computing*. Springer International Publishing. https://doi.org/10.1007/978-3-319-42999-1

Li, L., Yan, X., Peng, H., Xiu, Y., Gao, Y., & Wang, X. (2020). Diabetic retinopathy identification system based on transfer learning. *Journal of Physics: Conference Series*, *1544*, 12133. https://doi.org/10.1088/1742-6596/1544/1/012133

Li, Z., He, Y., Keel, S., Meng, W., Chang, R. T., & He, M. (2018). Efficacy of a Deep Learning System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus Photographs. *Ophthalmology*, *125*(8), 1199–1206. https://doi.org/10.1016/j.ophtha.2018.01.023

Lin, D. Y., Blumenkranz, M. S., Brothers, R. J., & Grosvenor, D. M. (2002). The sensitivity and specificity of single-field nonmydriatic monochromatic digital fundus photography with remote image interpretation for diabetic retinopathy screening: a comparison with ophthalmoscopy and standardized

mydriatic color photography. *American Journal of Ophthalmology*, *134*(2), 204–213. https://doi.org/10.1016/s0002-9394(02)01522-2

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Luo, G., Chutatape, O., & Krishnan, S. (2002). Detection and measurement of retinal vessels in fundus images using amplitude modified second-order Gaussian filter. *IEEE Transactions on Bio-Medical Engineering*, *49*, 168–172. https://doi.org/10.1109/10.979356

Majumder, S., & Kehtarnavaz, N. (2021). *Multitasking Deep Learning Model for Detection of Five Stages of Diabetic Retinopathy.* http://arxiv.org/abs/2103.04207

Martínez Rubio, M., Moya Moya, M., Bellot Bernabé, A., & Belmonte Martínez, J. (2012). Cribado de retinopatía diabética y teleoftalmología . In *Archivos de la Sociedad Española de Oftalmología* (Vol. 87, pp. 392–395). scieloes .

Masood, S., Luthra, T., Sundriyal, H., & Ahmed, M. (2017). Identification of diabetic retinopathy in eye images using transfer learning. *2017 International Conference on Computing, Communication and Automation (ICCCA)*, 1183–1187. https://doi.org/10.1109/CCAA.2017.8229977

Michelson, G. (2015). *Teleophthalmology in Preventive Medicine* (1st ed.). Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/978-3-662-44975-2

Mikla, V. I., & Mikla, V. V. (2014). *1 - Advances in Imaging from the First X-Ray Images* (V. I. Mikla & V. V. B. T.-M. I. T. Mikla (eds.); pp. 1–22). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-12-417021-6.00001-0

Mohammadian, S., Karsaz, A., & Roshan, Y. M. (2017). Comparative Study of Fine-Tuning of Pre-Trained Convolutional Neural Networks for Diabetic Retinopathy Screening. *2017 24th National and 2nd International Iranian Conference on Biomedical Engineering (ICBME)*, 1–6. https://doi.org/10.1109/ICBME.2017.8430269

Mompín Poblet, J. (1988). *Introducción a la bioingeniería.* Marcombo.

https://books.google.com.co/books?id=aqcaSGADoo4C

Mookiah, M R K, Acharya, U. R., Martis, R. J., Chua, C. K., Lim, C. M., Ng, E. Y. K., & Laude, A. (2013). Evolutionary Algorithm Based Classifier Parameter Tuning for Automatic Diabetic Retinopathy Grading: A Hybrid Feature Extraction Approach. *Know.-Based Syst.*, *39*, 9–22. https://doi.org/10.1016/j.knosys.2012.09.008

Mookiah, Muthu Rama Krishnan, Acharya, U. R., Chua, C. K., Lim, C. M., Ng, E. Y. K., & Laude, A. (2013). Computer-aided diagnosis of diabetic retinopathy: A review. *Computers in Biology and Medicine*, *43*(12), 2136–2155. https://doi.org/https://doi.org/10.1016/j.compbiomed.2013.10.007

Müller, R., Kornblith, S., & Hinton, G. (2019). *When Does Label Smoothing Help?* http://arxiv.org/abs/1906.02629

Murphey, Y. L., Guo, H., & Feldkamp, L. A. (2004). Neural Learning from Unbalanced Data. *Applied Intelligence*, *21*(2), 117–128. https://doi.org/10.1023/B:APIN.0000033632.42843.17

Organización Mundial de la Salud. (2020). *Informe mundial sobre la visión.*

Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Patel, R., & Chaware, A. (2020). Transfer Learning with Fine-Tuned MobileNetV2 for Diabetic Retinopathy. *2020 International Conference for Emerging Technology (INCET)*, 1–4. https://doi.org/10.1109/INCET49848.2020.9154014

Pham, H. N., Tan, R. J., Cai, Y. T., Mustafa, S., Yeo, N. C., Lim, H. J., Do, T. T. T., Nguyen, B. P., & Chua, M. C. H. (2020). Automated Grading in Diabetic Retinopathy Using Image Processing and Modified EfficientNet. In N. T. Nguyen, B. H. Hoang, C. P. Huynh, D. Hwang, B. Trawiński, & G. Vossen (Eds.), *Computational Collective Intelligence* (pp. 505–515). Springer International Publishing.

Pratt, H., Coenen, F., Broadbent, D. M., Harding, S. P., & Zheng, Y. (2016). Convolutional Neural Networks for Diabetic Retinopathy. *Procedia Computer Science*, *90*, 200–205. https://doi.org/https://doi.org/10.1016/j.procs.2016.07.014

Purves, D., Augustine, G. J., & Fitzpatrick, D. (2001). *Neuroscience* (2nd ed.). Sunderland (MA): Sinauer Associates.

https://www.ncbi.nlm.nih.gov/books/NBK10848/

Quellec, Gwenolé, Charrière, K., Boudi, Y., Cochener, B., & Lamard, M. (2017). Deep image mining for diabetic retinopathy screening. *Medical Image Analysis*, *39*, 178–193. https://doi.org/10.1016/j.media.2017.04.012

Quellec, Gwénolé, Lamard, M., Josselin, P. M., Cazuguel, G., Cochener, B., & Roux, C. (2008). Optimal wavelet transform for the detection of microaneurysms in retina photographs. *IEEE Transactions on Medical Imaging*, *27*(9), 1230–1241. https://doi.org/10.1109/TMI.2008.920619

Ranganathan, P., Pramesh, C., & Aggarwal, R. (2017). Common pitfalls in statistical analysis: Measures of agreement. *Perspectives in Clinical Research*, *8*(4), 187–191. https://doi.org/10.4103/picr.PICR_123_17

Real Academia Española. (2021). *Diccionario de la lengua española* (23rd ed.). https://dle.rae.es/retina

Reza, A., & Eswaran, C. (2011). A Decision Support System for Automatic Screening of Non-proliferative Diabetic Retinopathy. *Journal of Medical Systems*, *35*, 17–24. https://doi.org/10.1007/s10916-009-9337-y

Riordan-Eva, Paul; Cunningham, E. T. . (2012). *Vaughan y Asbury. Oftalmología general.* (18th ed.). McGRAW-HILL.

Rohan, T. E., Frost, C. D., & Wald, N. J. (1989). Prevention of blindness by screening for diabetic retinopathy: a quantitative assessment. *BMJ (Clinical Research Ed.)*, *299*(6709), 1198–1201. https://doi.org/10.1136/bmj.299.6709.1198

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. In *XXXVII Congreso Anual de La Sociedad Española de Ingeniería Biomédica (Caseib 2019)* (pp. 175–178).

Roychowdhury, S., Koozekanani, D. D., & Parhi, K. K. (2016). Automated detection of neovascularization for proliferative diabetic retinopathy screening. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, *2016*, 1300–1303. https://doi.org/10.1109/EMBC.2016.7590945

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2014). *ImageNet Large Scale Visual Recognition Challenge.*

http://arxiv.org/abs/1409.0575

Sahlsten, J., Jaskari, J., Kivinen, J., Turunen, L., Jaanio, E., Hietala, K., & Kaski, K. (2019). Deep Learning Fundus Image Analysis for Diabetic Retinopathy and Macular Edema Grading. *Scientific Reports*, *9*. https://doi.org/10.1038/s41598-019-47181-w

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. http://arxiv.org/abs/1801.04381

Savino, P. J. ., & Danesh-Meyer, H. V. . (2012). Color Atlas and Synopsis of Clinical Ophthalmology. Wills Eye Institute, Neuro-Ophthalmology (Wills Eye Institute Atlas Series), 2nd Edition. *Neuro-Ophthalmology*, *37*(5), 229. https://doi.org/10.3109/01658107.2013.824006

Saxena, G., Verma, D. K., Paraye, A., Rajan, A., & Rawat, A. (2020). Improved and robust deep learning agent for preliminary detection of diabetic retinopathy using public datasets. *Intelligence-Based Medicine*, *3–4*, 100022. https://doi.org/https://doi.org/10.1016/j.ibmed.2020.100022

Schmidhuber, J. (2014). *Deep Learning in Neural Networks: An Overview*. https://doi.org/10.1016/j.neunet.2014.09.003

Sengupta, S., Singh, A., Leopold, H. A., Gulati, T., & Lakshminarayanan, V. (2020). Ophthalmic diagnosis using deep learning with fundus images - A critical review. *Artificial Intelligence in Medicine*, *102*, 101758. https://doi.org/10.1016/j.artmed.2019.101758

Seoud, L., Chelbi, J., & Cheriet, F. (2015). *Automatic Grading of Diabetic Retinopathy on a Public Database*. https://doi.org/10.17077/omia.1032

Shaban, M., Ogur, Z., Mahmoud, A., Switala, A., Shalaby, A., Abu Khalifeh, H., Ghazal, M., Fraiwan, L., Giridharan, G., Sandhu, H., & El-Baz, A. S. (2020). A convolutional neural network for the screening and staging of diabetic retinopathy. *PLOS ONE*, *15*(6), 1–13. https://doi.org/10.1371/journal.pone.0233514

Shao, L., Zhu, F., & Li, X. (2015). Transfer Learning for Visual Categorization: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, *26*(5), 1019–1034. https://doi.org/10.1109/TNNLS.2014.2330900

Sheikh, S., & Qidwai, U. (2020). Using MobileNetV2 to Classify the Severity of Diabetic Retinopathy. *International Journal of Simulation Systems Science & Technology*. https://doi.org/10.5013/IJSSST.a.21.02.16

Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., & Summers, R. M. (2016). Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Transactions on Medical Imaging*, *35*(5), 1285–1298. https://doi.org/10.1109/TMI.2016.2528162

Sikder, N., Chowdhury, M., Arif, A., & Nahid, A. (2019). *Early Blindness Detection Based on Retinal Images Using Ensemble Learning*.

Sinha, T., Verma, B., & Haidar, A. (2017). Optimization of convolutional neural network parameters for image classification. *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1–7.

Skouta, A., Elmoufidi, A., Jai-Andaloussi, S., & Ochetto, O. (2021). Automated Binary Classification of Diabetic Retinopathy by Convolutional Neural Networks. In F. Saeed, T. Al-Hadhrami, F. Mohammed, & E. Mohammed (Eds.), *Advances on Smart and Soft Computing* (pp. 177–187). Springer Singapore.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*(56), 1929–1958. http://jmlr.org/papers/v15/srivastava14a.html

Stewart, M. (2017). *Diabetic Retinopathy: Current Pharmacologic Treatment and Emerging Strategies* (1st ed.). ADIS. https://doi.org/10.1007/978-981-10-3509-8

Suri, J. S. ., Campilho, A. ., Ng, E. Y. K. ., & Acharya, U. R. (2014). *Image Analysis and Modeling in Ophthalmology* (1st ed.). CRC Press. https://doi.org/https://doi.org/10.1201/b16510

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017). Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1). https://ojs.aaai.org/index.php/AAAI/article/view/11231

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2014). *Going Deeper with Convolutions*. https://arxiv.org/abs/1409.4842

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). *Rethinking the Inception Architecture for Computer Vision*. http://arxiv.org/abs/1512.00567

Ting, D. S. W., Cheung, C. Y.-L., Lim, G., Tan, G. S. W., Quang, N. D., Gan, A., Hamzah, H., Garcia-Franco, R., San Yeo, I. Y., Lee, S. Y., Wong, E. Y. M., Sabanayagam, C., Baskaran, M., Ibrahim, F., Tan, N. C., Finkelstein, E. A., Lamoureux, E. L., Wong, I. Y., Bressler, N. M., … Wong, T. Y. (2017). Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. *JAMA*, *318*(22), 2211–2223. https://doi.org/10.1001/jama.2017.18152

Tjoa, E., & Guan, C. (2020). A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 1–21. https://doi.org/10.1109/TNNLS.2020.3027314

Toennies, K. D. (2017). The Analysis of Medical Images. In *Guide to Medical Image Analysis: Methods and Algorithms* (pp. 1–22). Springer London. https://doi.org/10.1007/978-1-4471-7320-5_1

Tolias, Y. A., & Panas, S. M. (1998). A fuzzy vessel tracking algorithm for retinal images based on fuzzy clustering. *IEEE Transactions on Medical Imaging*, *17*(2), 263–273. https://doi.org/10.1109/42.700738

Too, E. C., Yujian, L., Njuki, S., & Yingchun, L. (2019). A comparative study of fine-tuning deep learning models for plant disease identification. *Computers and Electronics in Agriculture*, *161*, 272–279. https://doi.org/https://doi.org/10.1016/j.compag.2018.03.032

Tozer, K., Woodward, M. A., & Newman-Casey, P. A. (2015). Telemedicine and Diabetic Retinopathy: Review of Published Screening Programs. *Journal of Endocrinology and Diabetes*, *2*(4). https://doi.org/10.15226/2374-6890/2/4/00131

Tymchenko, B., Marchenko, P., & Spodarets, D. (2020). *Deep Learning Approach to Diabetic Retinopathy Detection*.

van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html

Van Der Maaten, L., Postma, E., & den Herik, J. (2009). Dimensionality reduction: a comparative review. *J Mach Learn Res*, *10*, 66–71.

Vieira, A., & Ribeiro, B. (2018). *Introduction to Deep Learning Business Applications for Developers*. Apress, Berkeley, CA. https://doi.org/https://doi.org/10.1007/978-1-4842-3453-2

Vilensky, J. A. ., Robertson, W. ., & Suarez-Quian, C. A. . (2015). *The Clinical Anatomy of the Cranial Nerves: The Nerves of "On Old Olympus Towering Top."* John Wiley & Sons.

Vlachos, M. (2009). Multi-scale retinal vessel segmentation using line tracking. *Computerized Medical Imaging and Graphics : The Official Journal of the Computerized Medical Imaging Society*, *34*, 213–227. https://doi.org/10.1016/j.compmedimag.2009.09.006

Wandishin, M. S., & Mullen, S. J. (2009). Multiclass ROC Analysis. *Weather and Forecasting*, *24*(2), 530–547. https://doi.org/10.1175/2008WAF2222119.1

Wang, W., Yang, Y., Wang, X., Wang, W., & Li, J. (2019). Development of convolutional neural network and its application in image classification: a survey. *Optical Engineering*, *58*(4), 1–19. https://doi.org/10.1117/1.OE.58.4.040901

Wilkinson, C. P., Ferris, F. L. 3rd, Klein, R. E., Lee, P. P., Agardh, C. D., Davis, M., Dills, D., Kampik, A., Pararajasegaram, R., & Verdaguer, J. T. (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

Xu, B., Wang, N., Chen, T., & Li, M. (2015). *Empirical Evaluation of Rectified Activations in Convolutional Network*. http://arxiv.org/abs/1505.00853

Yanoff, M., & Sassani, J. (2018). *Ocular Pathology* (8th ed.). Elsevier.

Zhang, W., Zhong, J., Yang, S., Gao, Z., Hu, J., Chen, Y., & Yi, Z. (2019). Automated identification and grading system of diabetic retinopathy using deep neural networks. *Knowledge-Based Systems*, *175*, 12–25. https://doi.org/https://doi.org/10.1016/j.knosys.2019.03.016

Zhao, H., Li, H., Maurer-Stroh, S., & Cheng, L. (2018). Synthesizing retinal and neuronal images with generative adversarial nets. *Medical Image Analysis*, *49*, 14–26. https://doi.org/https://doi.org/10.1016/j.media.2018.07.001

Zhou, K., Ruecker, D., & Fichtinger, G. (2019). *Handbook of Medical Image Computing and Computer Assisted Intervention*. Academic Press.

Zhu, J., Zhang, E., & Rio-Tsonis, K. (2012). *Eye Anatomy*. https://doi.org/10.1002/9780470015902.a0000108.pub2