Detección de enfermedades crónicas en imágenes de retina a través de técnicas de aprendizaje profundo. Mtro. Gendry Alfonso Francia 2023

Universidad Autónoma de Querétaro

Facultad de Ingeniería

Detección de enfermedades crónicas en imágenes de retina a través de técnicas de aprendizaje profundo.

Tesis

Que como parte de los requisitos para obtener el Grado de

Doctor en Ingeniería

Presenta

Mtro. Gendry Alfonso Francia

Dirigido por:

Dr. Saúl Tovar Arriaga

Querétaro, Qro, a 14 de diciembre de 2023

Detección de enfermedades crónicas en imágenes de retina a través de técnicas de aprendizaje profundo.

por

Gendry Alfonso Francia

Universidad Autónoma de Querétaro
Facultad de Ingeniería
Doctorado

Detección de enfermedades crónicas en imágenes de retina a través de técnicas de aprendizaje profundo.

Tesis

Que como parte de los requisitos para obtener el Grado de

Doctor en Ingeniería

Presenta

Mtro. Gendry Alfonso Francia

Dirigido por:
Dr. Saúl Tovar Arriaga

Dr. Saúl Tovar Arriaga
Presidente

Dr. Manuel Toledano Ayala
Secretario

Dr. Juvenal Rodríguez Reséndiz
Vocal

Dr. Jesús Carlos Pedraza Ortega
Suplente

Dr. Marco Antonio Aceves Fernández
Suplente

Centro Universitario, Querétaro, Qro.
Fecha de aprobación por el Consejo Universitario
14 de diciembre de 2023, México

"La inspiración existe, pero tiene que encontrarte trabajando."

DEDICATORIA

A mi madre y mi abuelo, que han sido esa antorcha que me va guiando en este camino de vida. A mis familiares y amigos que siempre han estado cuando los he necesitado.

RECONOCIMIENTO

En primer lugar, al Consejo Nacional de Ciencia y Tecnología (CONACyT), por proveer los fondos necesarios para realizar este trabajo de investigación. A la Universidad Autónoma de Querétaro y en específico a la Facultad de Ingeniería, por dar soporte y facilidades en mis estudios del posgrado. Reconocimiento especial al Dr. Saúl Tovar Arriaga, por su asesoría en la elaboración y desarrollo de esta tesis. A los profesores del programa doctoral y mención especial para todos los miembros de mi sínodo, que han sabido guiarme en este proceso.

RESUMEN

El *Deep Learning* (DL) se ha utilizado ampliamente para detectar anomalías en imágenes retinianas. Por lo general, esta tarea se ha centrado en un dominio específico, como las enfermedades relacionadas con el glaucoma o la retinopatía diabética, pero no ambas. En este estudio, proponemos identificar lesiones asociadas a ambas enfermedades utilizando un único modelo base, evitando el uso de múltiples modelos de DL. Se comenzó el estudio con un análisis comparativo del rendimiento de varios modelos de detección de objetos en la tarea de segmentar el disco y la copa ópticos. Los resultados arrojaron resultados excelentes y se seleccionó el modelo Cascade R-CNN. La tarea se complica por la necesidad de anotaciones en conjuntos de datos relacionados con daños en otro dominio para el que fue creado. Además, el tamaño y la forma de los objetos y el sesgo hacia las clases predominantes son evidentes. Varias técnicas caracterizan este trabajo, incluido el etiquetado suave para predicciones de máscaras, la distancia de Wasserstein normalizada para manejar objetos pequeños y experimentos en el muestreo de imágenes durante el entrenamiento con pérdida de entropía cruzada combinada con *Online Hard Negative Mining* o pérdida asimétrica. Para el refinamiento de resultados, el *cluster-weighted* con *Distance IoU* mejoró las predicciones finales. Basado en la precisión media promedio (mAP), una métrica estándar en modelos de detección de objetos, el resultado informado fue de 0.46. Cuatro conjuntos de datos públicos fueron empleados, REFUGE, ORIGA, G1020 y DDR. Se proporcionó un análisis de error detallado por categoría. En conclusión, se demostró la viabilidad de usar un solo modelo, mientras que las técnicas empleadas ayudaron a aumentar las métricas relacionadas con mAP. Nuestra investigación proporciona información novedosa sobre el uso de fotografías de retina para la predicción de biomarcadores sistémicos asociados con múltiples enfermedades.

Palabras claves: Glaucoma, Retinopatía Diabética, Detección de Objetos, Cascade R-CNN, Distancia Wasserstein, Pérdida Asimétrica, mAP

ABSTRACT

Deep learning (DL) has been widely used to detect abnormalities in retinal images. Typically, this task has been focused on a specific domain, such as diseases related to glaucoma or diabetic retinopathy, but not both. In this study, we propose to identify lesions associated with both diseases using a single base model, avoiding the use of multiple DL models. The study began with a comparative analysis of the performance of several object detection models on the task of segmenting the optic disc and cup. The results yielded excellent results and the Cascade R-CNN model was selected. The task is complicated by the need for annotations in datasets related to damage in another domain for which it was created. In addition, the size and shape of objects and bias towards predominant classes are evident. Several techniques characterize this work, including soft labeling for mask predictions, normalized Wasserstein distance for handling small objects, and experiments in image sampling during training with cross-entropy loss combined with Online Hard Negative Mining or asymmetric loss. For result refinement, cluster-weighted with Distance IoU improved final predictions. Based on mean average precision (mAP), a standard metric in object detection models, the reported result was 0.46. Four public datasets were employed, REFUGE, ORIGA, G1020, and DDR. A detailed error analysis by category was provided. In conclusion, the feasibility of using a single model was demonstrated, while the techniques employed helped to increase mAP-related metrics. Our research provides novel insights into the use of retinal photographs for the prediction of systemic biomarkers associated with multiple diseases.

Keywords: Glaucoma, Diabetic Retinopathy, Object Detection, Cascade R-CNN, Wasserstein Distance, Asymmetric Loss, mAP.

ÍNDICE DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ABREVIACIONES

ab. *Anchor box*
AP. *Average precision*
API. *Application Programming Interfaces*
APP. *Atrofia peripapilar*
ASL. *Asymmetric Loss*
CARAFE. *Content-Aware ReAssembly of Features*
CE. *Cross Entropy*
CNN. *Convolutional neural network*
CO. *Copa óptica*
COCO. *Common Objects in Context*
DL. *Deep learning*
DO. *Disco óptico*
EX. *Exudado duro*
FPN. *Feature Pyramid Network*
GAN. *Generative adversarial network*
GCNet. *Global Context Network*
gtb. *Ground truth box*
HE. *Hemorragia*
HTTP. *Hypertext Transfer Protocol*
IA. *Inteligencia Artificial*
IoU. *Intersección sobre unión*
ISNT. *Inferior > Superior > Nasal > Temporal*
M. *Multi-scale*
MA. *Microaneurisma*
MAILOR. *Mexican Advanced Imaging Laboratory for Ocular Research*
mAP. *Mean average precision*
MB. *Membrana de Brush*
ML. *Machine learning*
MS. *Multi-scale*
NMS. *Non maximum suppression*
NWD. *Normalize Wasserstein Distance*
OHEM. *Online Hard Example*
PAFPN. *Path Aggregation Feature Pyramid Network*
PR. *Precision-Recall*
RCD. *Relación Copa-Disco*
R-CNN. *CNN basada en regiones*
RD. *Retinopatía Diabética*
RDNP. *Retinopatía diabética no proliferativa*
RDP. *Retinopatía diabética proliferativa*
ReLU. *Rectified linear unit*
ResNet. *Red neuronal residual*
RKA. *Ranking Assigned*
RNA. *Red Neuronal Artificial*
ROI. *Region of interest*
RPN. *Region proposal network*
SE. *soft exudates*
SGD. *Stochastic gradient descend*
SSD. *Single Shot Multibox Detector*
TIDE. *Toolbox for Identifying Object Detection Errors*
WDIoUNMS. *Weighted Cluster-DIoU-NMS*
WM. *Without multi-scale*
YOLO. *You Only Look Once*

1. INTRODUCCIÓN

En esta sección se presenta el tema de investigación, su relevancia, así como los objetivos de esta. Elementos importantes que podrán ser encontrados son la motivación, la justificación de la investigación, el planteamiento del problema y la hipótesis. Una breve descripción del método de investigación fue agregada.

1.1. Motivación

A nivel mundial alrededor de 2200 millones de personas tienen deficiencia visual o ceguera, de ellas 1000 millones presentan una condición que pudo haberse evitado o que aún no ha sido tratada [1]. Una mejor comprensión de la magnitud de las necesidades de atención oftalmológica que actualmente se satisfacen en el sistema de salud es fundamental para una planificación eficaz.

Las enfermedades crónicas son un trastorno orgánico o funcional que obliga a una modificación del modo de vida del paciente y que persiste durante largo tiempo, entre ellas se encuentran las enfermedades cardíacas y la diabetes [2]. Enfermedades crónicas asociadas a la vista son la retinopatía (diabética e hipertensiva), la degeneración macular asociada a la edad, el glaucoma, entre otras, ver Figura 1-1.

Más allá de la prevención en el manejo de estas enfermedades es preciso dar un paso hacia delante enfocado en el análisis y detección de estas en fases tempranas de sus evoluciones, ya que un análisis es un estudio detallado, un examen cualitativo y cuantitativo de los componentes o sustancias del organismo según métodos especializados, con un fin diagnóstico [3]; mientras que detectar es la extracción de información particular de un flujo de información más grande sin cooperación específica o sincronización con el remitente. También se interpreta como la inspección y medición del objeto o fenómeno que no se puede observar directamente [4].

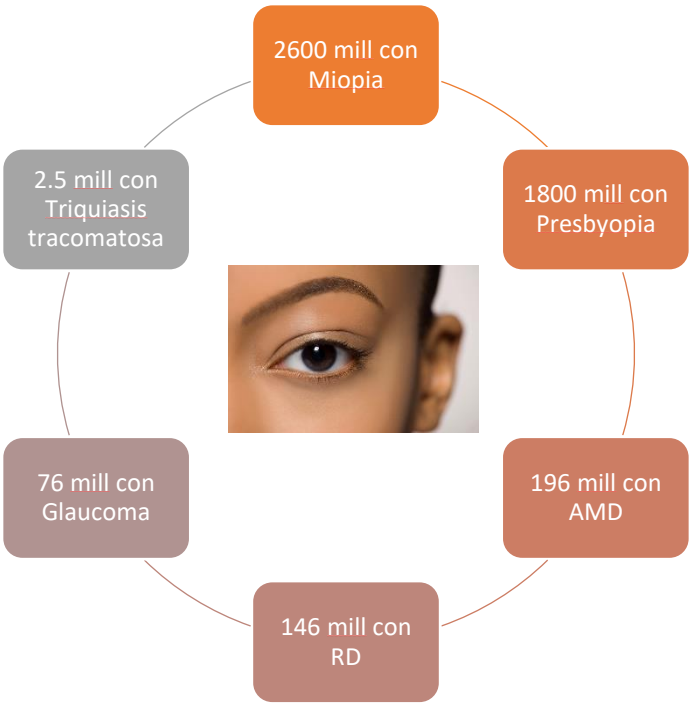


Figura 1-1: Prevalencia de enfermedades crónicas en la retina [1].

La retina es la capa de células nerviosas que recubren la pared posterior en el interior del ojo. Esta capa detecta la luz y envía señales al cerebro a través de unos elementos llamados conos y bastones para la formación de la imagen [5].

El empleo de imágenes de retina se justifica ya que es el único lugar del cuerpo humano donde se puede apreciar imágenes de las venas y arterias directamente, lo que nos brinda la oportunidad de apreciar la estructura y patología en vivo de la circulación humana [6].

Las imágenes médicas constituyen un nicho importante de información y diferentes técnicas de DL se han empleado para detectar patrones. El DL es una forma de representación del aprendizaje que utiliza múltiples pasos de transformación para identificar características muy complejas. Esta representación es jerárquica y permite al ordenador aprender conceptos complejos a partir de otros más simples. Los conceptos se construyen unos encimas de otros, quedando un gráfico de muchas capas, de ahí su definición de DL [7], [8].

Con esta investigación se diseñará una plataforma de software que permita la

detección de enfermedades crónicas en imágenes de retina haciendo uso de metodologías de procesamiento de dichas imágenes y técnicas de DL y que nos permita describir la causalidad de las detecciones en la toma de decisiones médicas.

1.2. La Retina

La vista es el sentido más fundamental que poseemos y la ceguera es quizás la mayor de todas las tragedias, ya que nos priva del sentido de la visión. Aunque todas las partes del ojo son importantes para percibir una buena imagen, la capa más vital para la visión es la retina. La retina es esencialmente una extensión del tejido cerebral, que recibe estimulación directa de la luz y las imágenes del mundo exterior.

En la retina se pueden apreciar estructuras como la red vascular, el disco óptico, donde confluyen las venas y arterias, así como la mácula y fóvea; además se pueden detectar biomarcadores de daño relacionados con alguna enfermedad crónica como microaneurismas (MA), hemorragias (HE), exudados duros (EX), etc., ver Figura 1-2.

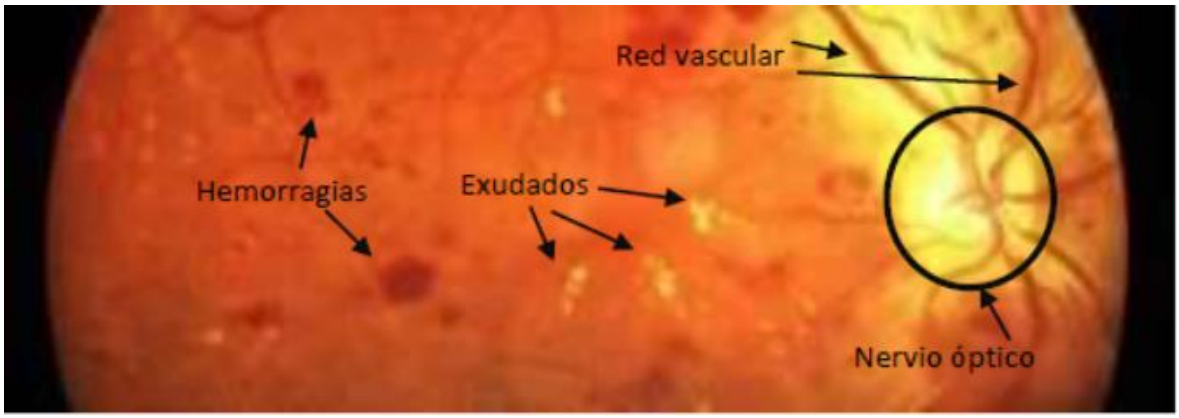


Figura 1-2: Imagen de retina (tomada y modificada de [9]*).*

Un biomarcador es definido como aquellas características biológicas, bioquímicas, antropométricas, fisiológicas, etc., objetivamente mesurables, capaces de identificar procesos fisiológicos o patológicos [10].

La retina humana es una organización delicada de neuronas, glía y vasos

sanguíneos nutritivos. En algunas enfermedades oculares, la retina se daña o se compromete, y se producen cambios degenerativos que eventualmente conducen a daños graves en las células nerviosas que transmiten los mensajes vitales sobre la imagen visual al cerebro [11]. Indicamos tres condiciones diferentes donde la retina está enferma y la ceguera puede ser el resultado final.

La degeneración macular relacionada con la edad es un problema retinal común del ojo envejecido y una de las principales causas de ceguera en el mundo. El área macular y la fóvea se comprometen debido a la degeneración del epitelio pigmentario detrás de la retina, que forma *drusen* (manchas blancas) y permite la fuga de líquido detrás de la fóvea. Los conos de la fóvea mueren, lo que causa una pérdida de visión central, por lo que no podemos leer ni ver detalles finos.

El glaucoma también es un problema común en el envejecimiento, donde la presión dentro del ojo aumenta. La presión aumenta porque la cámara anterior del ojo no puede intercambiar líquido correctamente por los métodos normales de flujo de salida acuoso. La presión dentro del vítreo aumenta y compromete los vasos sanguíneos de la cabeza del nervio óptico y eventualmente los axones de las células ganglionares, por lo que estas células vitales mueren. El tratamiento para reducir la presión intraocular es esencial en el glaucoma.

La retinopatía diabética es un efecto secundario de la diabetes que afecta la retina y puede causar ceguera. Los vasos sanguíneos vitales que nutren el ojo se comprometen, se distorsionan y se multiplican de manera incontrolable. El tratamiento con láser para detener la proliferación de vasos sanguíneos y la fuga de líquido hacia la retina es el tratamiento más común en la actualidad.

En esta investigación se abordarán biomarcadores de daño en la retina asociados al glaucoma y a la retinopatía diabética.

1.3. Justificación del problema

En 2021, la *International Diabetes Federation* estimó que uno de cada diez adultos (20–79 años) tiene diabetes (537 millones de personas), uno de cada dos adultos con diabetes no está diagnosticado (240 millones de personas) y el 9% del

gasto en salud mundial se destina a la diabetes (966,000 millones de dólares), entre otros datos relevantes [12]. En el mismo estudio se estima, que para el 2030 uno de cada nueve adultos tendrá diabetes (643 millones) y el gasto en salud relacionado con dicha enfermedad alcanzará 1 billón de dólares; así mismo para el año 2045 será un adulto cada ocho (783 millones) y el gasto asociado alcanzará los 1.1 billones de dólares.

La retinopatía diabética (RD) es una complicación de la diabetes que puede causar ceguera si no se trata a tiempo. La RD es la quinta causa principal de ceguera y deficiencia visual moderada a grave en adultos de 50 años o más. La prevalencia mundial estandarizada por edad de la ceguera debida a la RD ha aumentado del 14,9% al 18,5% entre 1990 y 2020 [13]. La RD puede desarrollarse incluso sin síntomas, por lo que es importante que las personas con diabetes se sometan a exámenes oftalmológicos regulares. Se sabe que la incidencia de la RD es de un 35% en pacientes diabéticos, que uno de cada tres pacientes con diabetes tendrán algún grado de RD, y uno de cada diez tendrán baja de visión por este motivo y que 145 millones de personas en el mundo sufren algún tipo de daño ocular por diabetes [14].

A nivel local, México ocupa el 5to lugar mundial de personas que viven con diabetes, con un total de 12 millones. De ese total un 5% puede padecer un edema macular (engrosamiento de la parte fina de la visión) con significativo deterioro de la visión central y hasta un 54% de las personas con diabetes tienen visión disminuida [14].

El glaucoma es una causa común de ceguera irreversible y está asociado con una pato-fisiología esencial que afecta a las células ganglionares de la retina, el estroma, los fotorreceptores, el cuerpo geniculado lateral y la corteza visual [15]. Se estima que esta enfermedad afecta a 80 millones de personas, de ellas, 1.5 millones son en México [16] y que hay hasta un 50% de personas que desconocen que presentan la enfermedad.

La tendencia del glaucoma a nivel global tampoco es muy poco esperanzadora.

De acuerdo a un estudio del 2006, para el año 2010 el número de personas con glaucoma sería de 60.5 millones, incrementándose a 79.6 millones de personas para el año 2020 y para el año 2040 este número se elevaría a 111 millones [17]. Solo en Estados Unidos el impacto económico asociado a la enfermedad es de 1500 millones de dólares.

Uno de los grandes desafíos para realizar el diagnóstico de las enfermedades antes mencionadas es la falta de cultura de la población para realizarse estudios periódicos. Incluso, personas diabéticas diagnosticadas, quienes deberían realizarse evaluaciones de retina periódicas, no las llevan a cabo por diferentes razones.

Por lo que se deriva la necesidad de un sistema automatizado, que proporcione una detección oportuna y confiable de enfermedades crónicas oculares, para mejorar la calidad de vida de las personas, la convivencia con las enfermedades y como posible influencia una reducción en los gastos económicos asociados al tratamiento.

Además, se busca una detección temprana, ya que puede permitir disponer de mejores indicadores económicos-financieros en los presupuestos de sanidad [18]. El sistema también puede proveer a médicos y especialistas de una herramienta de apoyo a los diagnósticos médicos; así como una plataforma para la experimentación con nuevas formas de detección en otras áreas de imágenes médicas.

1.4. Descripción del problema

En la actualidad, el incremento de la capacidad de cómputo y el desarrollo de nuevos algoritmos de IA, han propiciado el incremento de nuevas herramientas para el diagnóstico de enfermedades de la retina. Sin embargo, pocas han sido aprobadas para su uso comercial. Algunos ejemplos de softwares aprobados por la *Food and Drug Administration* son el sistema IDx-DR, para el diagnóstico de la RD, pero solo puede ser modificado por sus creadores y comercializado en USA, además que no emite un diagnóstico para otras posibles afecciones que tengan los

pacientes; el otro es el sistema *EyeArt AI Eye Screening System,* con permiso en la Unión Europea y Canadá, también para el diagnóstico de la RD [19]. Las capacidades de EyeArt recientemente fueron ampliadas, siendo capaz de detectar, adicionalmente Degeneración Macular asociada a la Edad y Glaucoma, así como utilizar diferentes equipos para la toma de imágenes [20]. Otros softwares han sido desarrollados o siguen en investigación, pero aún no son comerciales.

Otras limitantes actuales son la escasez de imágenes para el entrenamiento y validación de los modelos, mientras más cantidad mayor exactitud, precisión y sensibilidad de los modelos, contrarrestando también el problema de los falsos negativos que arrojan estos modelos, a pesar de los altos valores de exactitud de estos en la actualidad. Este tipo de errores son críticos para la atención de los pacientes.

Por otro lado, la no homogeneidad de los equipos de captura de las imágenes, con diferentes calidades y resoluciones constituyen un obstáculo para los algoritmos de DL.

También nos encontramos con sistemas con poca capacidad explicadora para los profesionales de la salud, lo que los hace en muchos casos rechazar el uso de las nuevas tecnologías. Este problema es conocido como el fenómeno de “caja negra”, ya que el resultado obtenido está basado en un entrenamiento y aprendizaje intensivo [21].

Siguiendo la tendencia mundial en el diagnóstico por telemedicina, el Departamento de Retina y Vítreo del Instituto Mexicano de Oftalmología a través del *Mexican Advanced Imaging Laboratory for Ocular Research* (MAILOR) creó un sistema de software de telemedicina para dar diagnóstico a pacientes de manera remota. La adquisición de imágenes se realiza por medio de cámaras portátiles no midriáticas, en 25 unidades localizadas en toda la República Mexicana. Las imágenes se envían por internet mediante el software especializado, desarrollado por el MAILOR y este es analizado en el laboratorio por médicos certificados. El resultado de la evaluación es enviado en máximo 48 horas. En el 2019, se realizó

el escrutinio de alrededor de 12,000 pacientes [9].

A partir de lo analizado se busca realizar un sistema automatizado que no solo nos dé una mayor exactitud en los análisis sobre imágenes de retina, sino que nos dé una retroalimentación instantánea de una decisión clínica en tiempo real y un resultado consistente sin importar edad, raza y etnicidad. Además de que provea una capacidad explicadora sobre las predicciones hechas en las imágenes adquiridas y normalizadas en los centros de salud nacionales.

La incógnita científica de esta propuesta de investigación es determinar si la detección de múltiples biomarcadores de daño en imágenes de retina, a través de su preprocesamiento y posterior análisis a través de técnicas de DL, generará patrones y predicciones de enfermedades crónicas en la retina que deriven en clasificaciones y aprendizajes.

1.5. Hipótesis

La detección y localización de diferentes anomalías en imágenes de retina de fondo de ojo, a través del ajuste de modelos de detección de objetos, se logrará con precisiones equiparables a algoritmos de DL especializados, mientras proporciona una base a las decisiones del usuario de manera interpretable.

1.6. Objetivo general

Desarrollar un marco de trabajo integral que permita la detección de patrones de riesgo de enfermedades crónicas en imágenes de retina a través de algoritmos de DL.

1.7. Objetivos específicos

- Diseñar y desarrollar algoritmos de DL para clasificar y detectar enfermedades crónicas en imágenes de retina.

- Diseñar un marco de trabajo que permita integración de modelos de redes neuronales.

- Desarrollar la integración entre herramienta de software y marco de trabajo

para la visualización y evaluación de enfermedades crónicas en imágenes de retina.

- Determinar el grado de aproximación de los modelos propuestos contra otras técnicas de DL y por un equipo de médicos especializados.

1.8. Estrategia de investigación

La investigación exploratoria es la estrategia adecuada para este estudio, ya que se trata de un tema poco conocido y que requiere una comprensión más profunda. El objetivo de este tipo de investigación es familiarizar al investigador con el tema, la situación actual y los métodos y técnicas utilizados. Las fuentes de información más apropiadas son la bibliografía existente y las entrevistas a personas vinculadas a la problemática.

1.9. Métodos científicos

En esta investigación se utilizaron los siguientes métodos científicos:

- Histórico-lógico: Este método permitió analizar el desarrollo histórico de la detección de lesiones en imágenes de retina, identificando deficiencias y proponiendo soluciones.
- Hipotético-deductivo: A partir de una hipótesis, se dedujeron nuevos conocimientos y predicciones, que se verificaron mediante experimentos.
- Analítico-sintético: Se analizó la teoría existente sobre el tema, para aplicarla al diseño de un algoritmo y adquirir una mayor comprensión del fenómeno.
- Modelación: Se creó un modelo matemático del fenómeno, que se asemeja al objeto real mediante una abstracción.

1.10. Estructura de la tesis

El presente documento cuenta con 3 capítulos:

Capítulo 1 - Marco Teórico: El marco teórico de la investigación describe los conceptos fundamentales del dominio del problema y el objeto de estudio. También analiza la situación actual del problema y revisa el estado del arte de trabajos similares. Finalmente, presenta la fundamentación de las tecnologías utilizadas

para el diseño del sistema y las propuestas para su implementación y desarrollo.

Capítulo 2 - Metodología: Se realiza un análisis de los principales procesos vinculados al objeto de estudio y al campo de acción del trabajo. Se pasa al diseño del modelo a desarrollar teniendo en cuenta los resultados del análisis de los procesos vinculados al objeto de estudio. Se implementa el modelo mediante su codificación.

Capítulo 3 - Resultados y Evaluación: En este apartado son mostrados los resultados tras implementar la metodología propuesta. Tablas numéricas y figuras descriptivas son proveídas, a la vez que se establece una comparación con el estado del arte investigado previamente. Finalmente, un extenso sub-epígrafe de discusión de resultados nos da evidencias claras de la aportación de la investigación.

Conclusiones - Recomendaciones y trabajos futuros: Por último, se abordan las conclusiones de la investigación, el aporte, limitaciones encontradas y posibles líneas de investigación identificadas.

2. MARCO TEÓRICO

En este capítulo se aborda la fundamentación teórica de la investigación a través de plantear los elementos básicos asociados a los conceptos médicos y las bases y principios de la tecnología utilizada. Una descripción de las herramientas y materiales utilizados fue proveída para que una reproducción de la investigación se pueda llevar a cabo. El estado del arte fue analizado, identificando las fortalezas actuales, así como las debilidades que dan pie a la propuesta de solución de la investigación.

2.1. Revisión médica

2.1.1. Glaucoma

Es difícil definir el glaucoma con precisión, en parte porque el término engloba un grupo diverso de trastornos. Todos los glaucomas tienen un común una típica

neuropatía óptica potencialmente progresiva, que se asocia a pérdida de campo visual a medida que avanza la lesión, y en la que la presión intraocular es un factor modificable fundamental [22].

De acuerdo con la etiología del glaucoma, este se clasifica: Glaucoma Primario, Glaucoma Congénito, Glaucoma Secundario y Glaucoma Absoluto; todas las clasificaciones con subcategorías, pero el glaucoma primario de ángulo abierto es el más común de todos. El principal mecanismo de pérdida visual en el glaucoma es la apoptosis de las células ganglionares de la retina, que son las células que transmiten la información visual al cerebro. Esta apoptosis conduce al adelgazamiento de las capas nuclear interna y de fibras nerviosas de la retina, así como a la pérdida de axones en el nervio óptico. El disco óptico, que es el punto de salida del nervio óptico del ojo, se vuelve atrófico, con agrandamiento de la copa óptica. En todos los pacientes con glaucoma, la necesidad de tratamiento y su eficacia se evalúan mediante la determinación periódica de la presión intraocular (tonometría), inspección de los discos ópticos y medición de los campos visuales [23].

En el diagnóstico del glaucoma, los especialistas se centran en la cabeza del nervio óptico, donde se evalúa el anillo neurorretiniano, que es el tejido anaranjado-rosado situado entre el límite externo de la excavación y el borde de la papila óptica, también conocida como copa óptica, analizando la regla ISNT, donde, dónde es más ancha la parte inferior, seguido por las zonas superior, nasal y temporal. También se analiza la razón entre el disco y la copa óptica, normalmente el cociente vertical, más que el horizontal. Las copas con diámetro pequeño tienen excavaciones pequeñas y viceversa. Cualquier persona que presente una razón entre el disco y la copa óptica mayor a 0.2 se considera sospechosa, ver Figura 2-1.

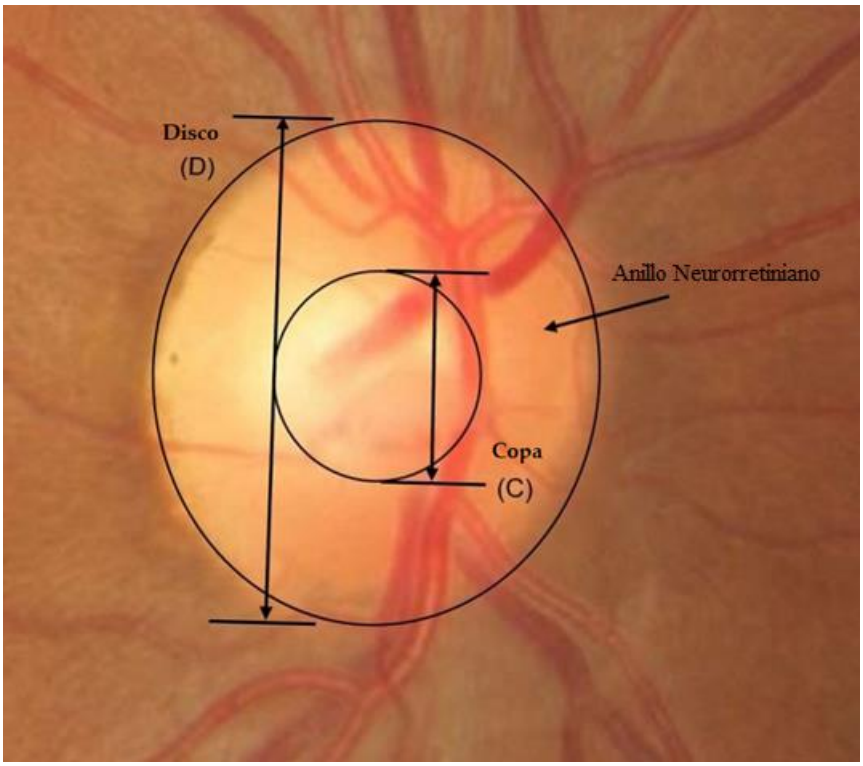


Figura 2-1: Nervio óptico sin presencia de glaucoma. Se identifican el disco y copa ópticas, así como el anillo neurorretiniano.

2.1.1.1. Copa y disco ópticos

El disco óptico normal es redondo y de color rosa. En el centro del disco óptico hay una depresión, llamada copa fisiológica. El tamaño de la copa fisiológica varía de persona a persona, pero en general es más pequeña en los ojos hipermétropes y más grande en los ojos miopes.

El glaucoma es una enfermedad que daña el nervio óptico. El primer signo de glaucoma es el adelgazamiento de la capa de fibras nerviosas retinianas en la región que rodea el disco óptico. La atrofia óptica glaucomatosa produce cambios específicos en el disco caracterizados principalmente por la pérdida de tejido del disco, que se manifiesta como agrandamiento de la copa del disco óptico y palidez en el área de la excavación. La Figura 2-2 muestra cambios comunes en un ojo glaucomatoso.

Figura 2-2: Excavación glaucomatosa ("en forma de olla") del disco óptico con desplazamiento nasal de los vasos retinianos y apariencia completamente ahuecada del disco óptico [23].

La relación de diámetro entre la copa y disco ópticos (RCD), se ha utilizado tradicionalmente para describir la cabeza del nervio óptico, así como para semicuantificar el glaucoma que puede presentar un paciente. En ojos normales, el RCD horizontal es mayor que el vertical, ya que la copa tiene una forma ovalada horizontal y el disco una forma ovalada vertical. Por lo tanto, el anillo neurorretiniano es más ancho en las áreas inferiores y superiores que las temporales y nasales. Dado que el glaucoma afecta en sus primeras y medias etapas avanzadas de forma preferente las regiones inferior y superior del disco óptico, la RCD vertical aumenta más que la RCD horizontal en los ojos con glaucoma progresivo. Esto es un indicador de que la relación de RCD vertical es más importante que la horizontal en el diagnóstico del glaucoma [24].

La RCD es una medida independiente de la magnificación, lo que significa que no se ve afectada por el tamaño del ojo o por la potencia de las lentes utilizadas para examinarlo. Esto hace que la RCD sea una medida muy útil para evaluar el daño al nervio óptico en pacientes con glaucoma. La mayoría de los métodos para medir la RCD no requieren el uso de ningún dispositivo especial, por lo que se pueden realizar en cualquier consultorio oftalmológico. Esto hace que la RCD sea una medida muy accesible y práctica para el seguimiento de los pacientes con glaucoma.

2.1.1.2. Atrofia peripapilar (Alfa y Beta)

Recientes estudios han demostrado en cierto grado la relación entre algunas regiones peripapilares y el glaucoma, así como un indicador de lesión precoz en pacientes con hipertensión ocular. La atrofia peripapilar (APP) se localiza alrededor de la cabeza del nervio óptico y cuatro zonas se pueden diferenciar, la Alfa, la Beta, la Gamma y la Delta. Las más comunes son las dos primeras, donde centraremos este estudio [22].

La atrofia Alfa se caracteriza por alteraciones del epitelio pigmentario retiniano superficial y la presencia de la Membrana de Brush (MB). Es la más periférica de todas las zonas y está presente en casi todos los ojos normales y se estima que tiende a ser más grande y frecuente en ojos glaucomatosos [22], [24].

La atrofia Beta está definida por la presencia de la MB y la ausencia del epitelio pigmentario retiniano. La aparición y el tamaño de la zona beta se correlacionan con la pérdida glaucomatosa del borde neurorretinal dentro del disco óptico, la pérdida glaucomatosa del campo visual, la disminución del diámetro de las arterias retinianas en los ojos con glaucoma y la disminución del diámetro de la parte retrobulbar del nervio óptico medida por sonografía [24].

Una zona Beta grande, se asocia a menudo con las siguientes características:

- Excavación glaucomatosa poco profunda del disco.
- Baja frecuencia de hemorragias del disco.
- Defectos localizados detectables de la capa de fibras nerviosas retinianas.
- Pérdida mayormente concéntrica del borde neurorretinal.
- Mediciones de la presión intraocular normales o casi normales.

Tanto la zona Alfa como la zona Beta son significativamente mayores y esta última ocurre con mayor frecuencia en ojos con atrofia glaucomatosa del nervio óptico en comparación a ojos normales. En la Figura 2-3 pueden apreciar la presencia de ambas atrofias.

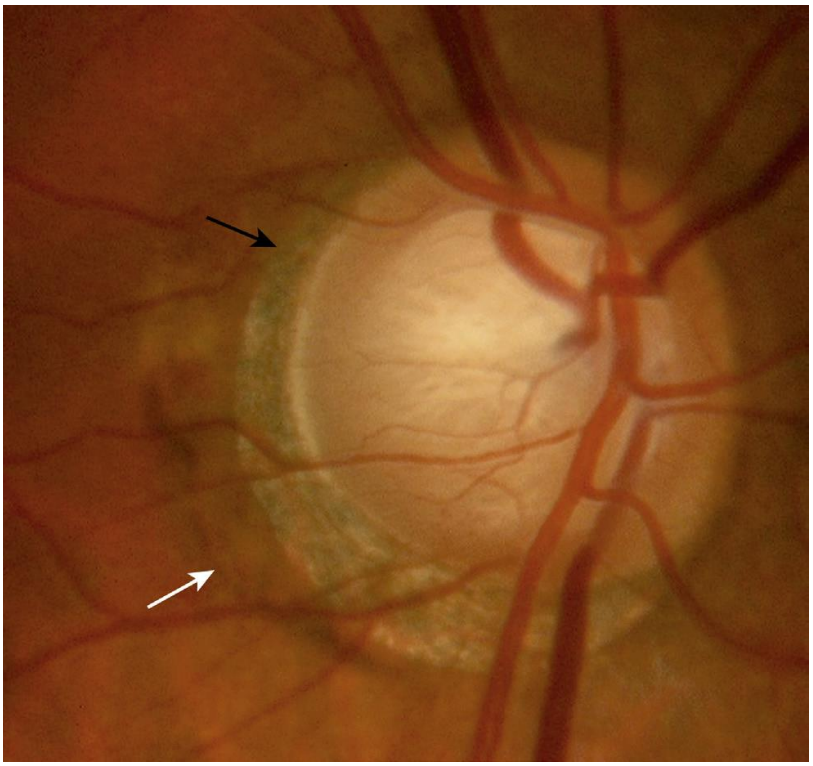


Figura 2-3: Presencia de atrofias peripapilares. Alfa (flecha blanca), Beta (flecha negra) [22].

2.1.2. Retinopatía diabética

La RD se caracteriza por daños en los vasos sanguíneos de la retina que se vuelven permeables o se bloquean. La pérdida de visión se produce con mayor frecuencia debido a la inflamación de la parte central de la retina, lo que puede provocar un deterioro de la visión. También pueden crecer vasos sanguíneos anormales de la retina, que pueden sangrar o causar cicatrices en la retina y ceguera [1].

La RD es una microangiopatía progresiva. La hiperglucemia crónica conduce a una respuesta metabólica mediada por un aumento de los productos finales de glicación avanzada, polioles, especies reactivas de oxígeno, eicosanoides, óxidos nítricos y moléculas de adhesión intercelular, y por la activación de la vía de la proteína kinasa C, lo que produce un daño endotelial microvascular, leucoestasis en los capilares retinianos y oclusión capilar. La isquemia retiniana interna resultante desencadena el crecimiento de nuevos vasos sanguíneos, que a su vez produce la ruptura de la barrera retiniana interna de la sangre y la fuga vascular

[23].

La clasificación empleada en el *Early Treatment Diabetic Retinopathy Study* es muy utilizada en el mundo y engloba las siguientes categorías:

- Retinopatía diabética no proliferativa (RDNP).
  - Ausencia de RD.
  - RDNP muy leve.
  - RDNP leve.
  - RDNP moderada.
  - RDNP grave.
  - RDNP muy grave.
- Retinopatía diabética proliferativa (RDP).
  - RDP leve-moderada.
  - RDP de alto riesgo.
  - Oftalmopatía diabética avanzada.

Normalmente se asocia a la RD la presencia de HE, MA, EX y exudados blandos (*soft exudates*, SE), los cuales se describen a continuación [22]. Ver Figura 2-4 para una identificación gráfica.

2.1.2.1. Hemorragias

Se presentan en una variedad que engloban hemorragias de la capa de fibras nerviosas retinianas y se forman a partir de las arteriolas precapilares superficiales, hemorragias intrarretinianas que proceden del extremo venoso de los capilares y se localizan en las capas intermedias más compactas, y las hemorragias redondas oscuras más profundas que representan infartos retinianos.

2.1.2.2. Microaneurismas

Son evaginaciones de la pared capilar que pueden formarse por dilatación focal de zonas con ausencia de pericitos o por fusión de dos ramas de un asa capilar. Los MA pueden rezumar elementos del plasma hacia la retina debido a la alteración de la barrera hematorretiniana o bien trombosarse. Suelen ser el signo más precoz de RD. Se identifican a menudo como puntos rojos.

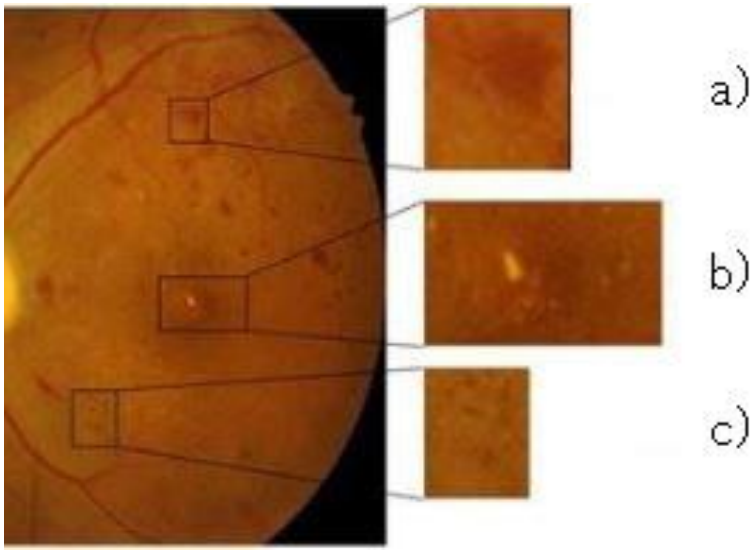


Figura 2-4: Imagen de retina con lesiones asociadas a la RD. a) Hemorragia, b) Exudados, c) Micro-aneurismas [25].

2.1.2.3. Exudados duros

Se deben a edemas retinianos y se forman en la unión de la retina normal y edematosa. Se componen de lipoproteínas y macrófagos llenos de lípidos y se identifican como lesiones céreas amarillentas, a menudo alrededor de MA y con el tiempo tienden a aumentar de número y tamaño.

2.1.2.4. Exudados blandos (SE)

Están formados por acumulaciones de residuos neuronales dentro de la capa de fibras nerviosas y se producen como resultado de la destrucción isquémica de los axones nerviosos. Se identifican como pequeñas lesiones superficiales blanquecinas y plumosas.

2.2. Revisión tecnológica

En el siguiente epígrafe estaremos revisando le núcleo de la tecnología utilizada para abordar el problema antes mencionado. Se propone utilizar el DL, una forma de inteligencia artificial que analiza datos para identificar patrones y tendencias. En las imágenes médicas, el DL puede utilizarse para predecir enfermedades o detectar anomalías. Específicamente en la detección de anomalías, avances recientes han introducido los algoritmos de detección de objetos, los cuales serán la técnica primaria empleada en esta investigación.

2.2.1. Redes neuronales

Una red neuronal artificial (RNA) es básicamente un modelo matemático para procesamiento de información, implementado por software o por hardware y que tiene una entidad propia expresada en un conjunto de parámetros internos, arquitecturas y modelos que la diferencian de otras técnicas y que reúne algunas características comunes con las redes neuronales biológicas [26], [27].

Una RNA es un sistema que se compone de una serie de neuronas, cada una de las cuales tiene una función de activación. La función de activación determina cómo se procesan las entradas de la neurona. Las entradas de una neurona se multiplican por sus pesos y luego se suman a un sesgo. Esta suma representa una combinación lineal de las entradas y sus pesos. La función de activación se aplica a esta suma para agregar un elemento de no linealidad. La representación matemática es la siguiente:

$$\partial = \sigma(W^T x + b) \qquad (1)$$

Donde W es la matriz de pesos, x las unidades de entrada y b el bias, σ representa la función de transferencia que tradicionalmente son las sigmoides cuya ecuación es:

$$f(x) = \frac{1}{1+e^{-x}} \qquad (2)$$

y la tangente hiperbólica (Brio and Molina 2006):

$$f(x) = \frac{1-e^{-2x}}{1+e^{-2x}} \qquad (3)$$

El esquema tradicional de una red neural se muestra en la Figura 7, donde se aprecia una capa de entrada, una capa oculta y una capa de salida.

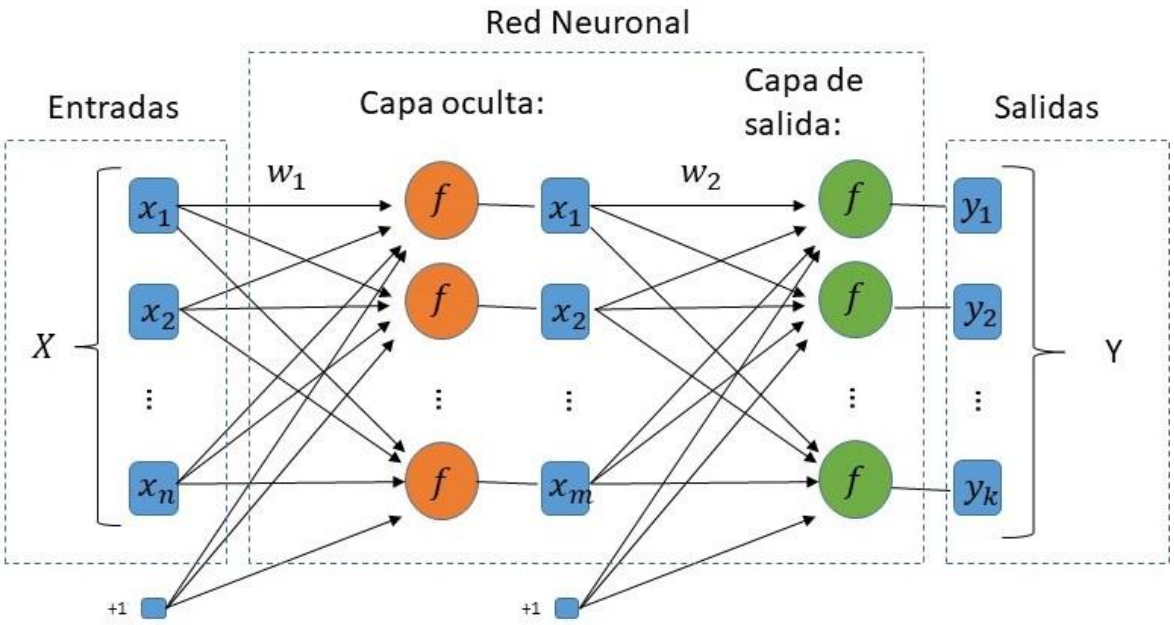


Figura 2-5: Topología de una red neuronal con una capa oculta.

Donde:

f: función de transferencia que se aplica.

$(x_1, \dots, x_n)$: vector extendido de la capa de entrada.

$(x_1, \dots, x_m)$: vector extendido de la capa oculta.

$(y_1, \dots, y_k)$: vector extendido de la capa de salida, que se aproxima al valor deseado o esperado.

2.2.2. Redes neuronales profundas

Las RNA son el tipo de algoritmo de aprendizaje que constituyen la base de los métodos de DL. Se consideran profundas cuando tienen múltiples capas ocultas y van creciendo en complejidad y abstracción.

El poder del DL proviene de la capacidad de las RNA para aprender patrones complejos en los datos. Esto se logra mediante la composición de múltiples funciones no lineales. Cada función no lineal aprende a transformar los datos de una manera específica. Al componer estas funciones, las RNA pueden aprender patrones cada vez más complejos [29].

Para el trabajo con redes neuronales profundas hasta la actualidad se ha comprobado que la función de activación, unidad lineal rectificada (*ReLU*, por sus siglas en inglés) ha brindado el mejor desempeño, ya que ha obtenido resultados de menor error comparada con las funciones logísticas antes mencionadas [30]. La ecuación 4 muestra su definición y para una representación gráfica, ver Figura 2-6.

$$f(x) = \max(0, x) \qquad (4)$$

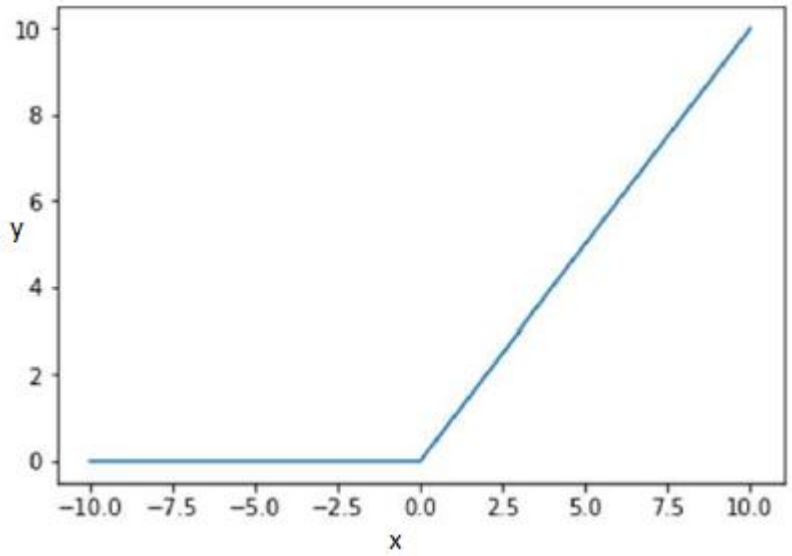


Figura 2-6: Gráfica de función ReLU [31].

Sin embargo, si se utilizan demasiadas funciones, la red puede sobreajustarse a los datos de entrenamiento. El sobreajuste ocurre cuando la red aprende los datos de entrenamiento tan bien que no puede generalizar a nuevos datos. Esto significa que la red no podrá realizar predicciones precisas para datos que no ha visto antes.

Hay varias formas de evitar el sobreajuste en las RNA. Una forma es utilizar una RNA con menos parámetros. Esto se puede lograr utilizando funciones más eficientes o reduciendo el número de capas en la red. Otra forma de evitar el sobreajuste es utilizar técnicas de regularización, como la deserción neuronal o la normalización por lotes [32].

Las redes neuronales profundas fueron consideradas difíciles de entrenar hasta que, en 2006, Bengio, Hilton y Salakhudinov desarrollaron un método de

entrenamiento capa a capa no supervisado seguido de un ajuste supervisado sobre una red apilada. Este método mostró un buen rendimiento, y las redes neuronales profundas se han convertido en una herramienta popular en el procesamiento de imágenes médicas. Las redes neuronales convolucionales (CNN, por sus siglas en inglés) son una de las arquitecturas más populares, y se describirán a continuación. Otras arquitecturas también usadas en imágenes médicas son las *auto-encoder*, las máquinas de Boltzmann restringidas, redes neuronales convolutivas *multi-stream* y U-Net [27].

Las computadoras leen las imágenes como píxeles y se expresan como una matriz de NxNx3. La capa convolucional utiliza un conjunto de filtros para detectar la presencia de características o patrones específicos presentes en la imagen original de entrada, el filtro se desliza por la imagen y se calcula un producto de puntos para obtener un mapa de activación, posteriormente se aplica la función de activación ReLU antes comentada para romper la linealidad de la imagen; después se pasa a la capa de *pooling*, la que se encarga de reducir la cantidad de parámetros y el cálculo de la red, controlando el sobreajuste y reduciendo progresivamente el tamaño espacial de la red, la técnica utilizada aquí comúnmente es *max-pooling*, el paso posterior es introducir los parámetros dentro de una capa completamente conectada, siendo el último paso [33]. La Figura 2-7 muestra la breve descripción antes expuesta.

Independientemente del diseño o arquitectura que se emplee, algo que definitivamente a contribuido al auge de estas técnicas ha sido la disponibilidad de GPU y librerías relacionadas con estos como CUDA y OpenCL. Junto al avance del hardware han aparecido diferentes paquetes de software libre que proveen eficiencia al diseño de nuestras RNA. Los paquetes más populares son Caffe, Tensorflow, Theano y Torch [27].

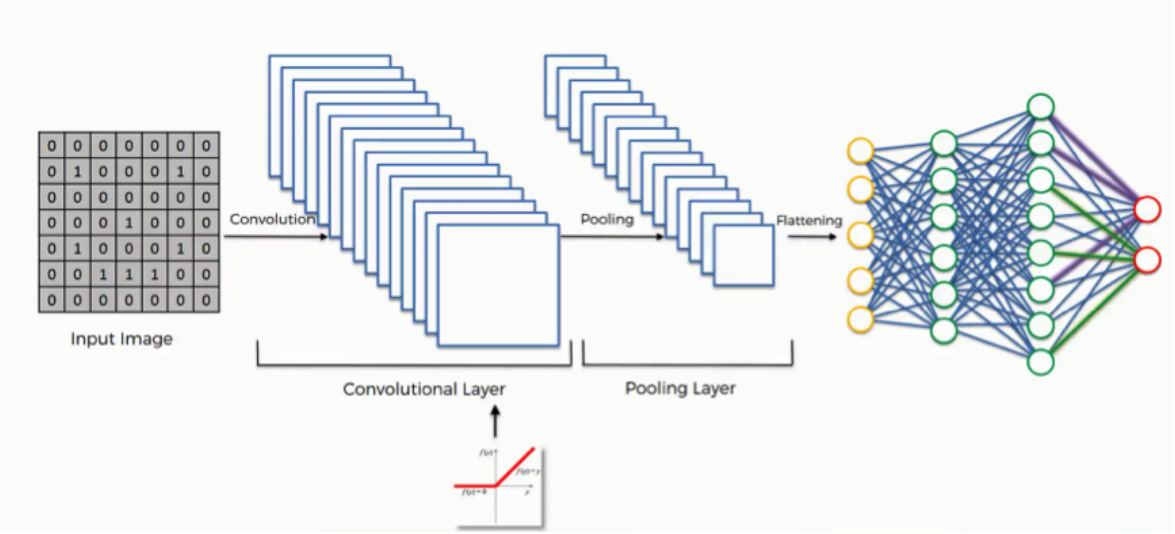


Figura 2-7: Ejemplo de arquitectura de red neuronal convolutiva (tomado de [34]).

El uso de CNN se justifica debido a su modelo, donde la red compuesta por neuronas convolucionales extrae características de las imágenes suministradas y reduce su dimensionalidad, comportamiento parecido al de la corteza visual primaria de nuestro cerebro biológico [33].

La profundidad y la anchura de esta capacidad pueden controlarse para obtener mejores suposiciones sobre la naturaleza de las imágenes. Por ejemplo, una profundidad mayor permite modelar relaciones a mayor distancia, mientras que una anchura mayor permite modelar relaciones entre más píxeles. Por lo tanto, en comparación con las redes neuronales de retroalimentación estándar con capas de tamaño similar, las CNN tienen muchas menos conexiones y parámetros, siendo más fáciles de entrenar, mientras que su rendimiento teóricamente óptimo probablemente sea sólo ligeramente peor [35].

2.2.3. Transfer Learning

El entrenamiento de una CNN desde cero puede ser muy exigente en términos de tiempo y recursos computacionales, especialmente si el conjunto de datos es grande. El área médica es un ejemplo, donde las imágenes juegan un rol importante para el diagnóstico de enfermedades. Dichas imágenes son generadas por equipos médicos especializados y el etiquetado a menudo es realizado por médicos especializados, siendo la tarea costosa y en la mayoría de los casos es difícil

colectar suficientes datos para el entrenamiento. Por esta razón, es común pre-entrenar una CNN en un conjunto de datos muy grande, como ImageNet [36], y luego utilizar esta CNN preentrenada como punto de partida para entrenar una CNN para una tarea específica [37].

Los tres grandes escenarios del aprendizaje por transferencia son los siguientes:

- Red convolucional como extractor de características fijo: para usarlo se deben seguir los siguientes pasos:
  - Eliminar la última capa totalmente conectada de la red convolucional.
  - Extraer las características o códigos de la CNN para todas las imágenes del nuevo conjunto de datos.
  - Entrenar un clasificador lineal para el nuevo conjunto de datos utilizando los códigos CNN como entrada.
- Ajuste fino de la red convolucional: es una técnica que permite mejorar el rendimiento de la CNN en una tarea específica con menos datos y recursos computacionales. Para realizar el ajuste fino hacer los siguientes pasos:
  - Reemplazar la última capa totalmente conectada de la CNN preentrenada por un nuevo clasificador específico para la nueva tarea.
  - Entrenar el nuevo clasificador utilizando los códigos CNN de las imágenes del nuevo conjunto de datos.
  - Ajustar los pesos de las capas de la CNN preentrenada utilizando la retropropagación. Es posible ajustar todos los niveles de la CNN, o es posible mantener algunos de los niveles anteriores fijos (debido a preocupaciones de sobreajuste) y sólo ajustar una parte de alto nivel de la red.
- Modelos preentrenados: la red se reentrena completamente, pero a partir de pesos que ya han sido entrenados en otro conjunto de imágenes. Esto reduce el número de iteraciones necesarias para alcanzar el nivel de

precisión deseado.

Para decidir qué tipo de aprendizaje por transferencia utilizar depende de dos factores fundamentales, el tamaño del nuevo conjunto de datos y su similitud con el conjunto de datos original. Teniendo en cuenta que las características de las CNN son más genéricas en las primeras capas y más específicas del conjunto de datos original en las capas posteriores, aquí hay algunas reglas generales comunes para navegar por los 4 escenarios principales [38]:

- El nuevo conjunto de datos es pequeño y similar al conjunto de datos original.
- El nuevo conjunto de datos es grande y similar al conjunto de datos original.
- El nuevo conjunto de datos es pequeño pero muy diferente del conjunto de datos original.
- El nuevo conjunto de datos es grande y muy diferente del conjunto de datos original.

2.2.4. Modelos de detección de objetos

La detección de objetos tiene por objetivo contar los objetos de una escena y determinar y rastrear sus ubicaciones precisas, etiquetándolos con exactitud. Estos modelos son más resistentes a las oclusiones, iluminación difícil y escenas complejas.

En la actualidad, los modelos de detección de objetos se utilizan para la estimación de la pose, la detección de vehículos y la vigilancia, entre otras aplicaciones. Estos algoritmos intentan dibujar una caja delimitadora alrededor del objeto de interés. No tiene por qué ser necesariamente una, pueden ser varias las dimensiones de la caja y diferentes los objetos.

Con el desarrollo tecnológico, el auge del DL y la irrupción de las CNN, los modelos de detección de objetos se han agrupado en dos categorías: detección de un estado y detección de dos estados, donde el primero es un proceso de detección en un solo paso, mientras que el segundo es un proceso de refinamiento en dos

pasos.

Dentro de los modelos de un solo estado se encuentran YOLO (*You Only Look Once*), SSD (*Single Shot Multibox Detector*) y RetinaNet. Estos modelos son simples y rápidos, pero típicamente tienen menor precisión. Por otro lado, los modelos de dos estados primero generan un conjunto de cuadros delimitantes, para luego clasificar el objeto y refinar la detección, permitiendo estas características un rendimiento superior en cuanto a precisión. Debido a esto, esta investigación se centró en los de dos estados, en la Figura 2-8 se puede ver una presentación de los modelos utilizados.

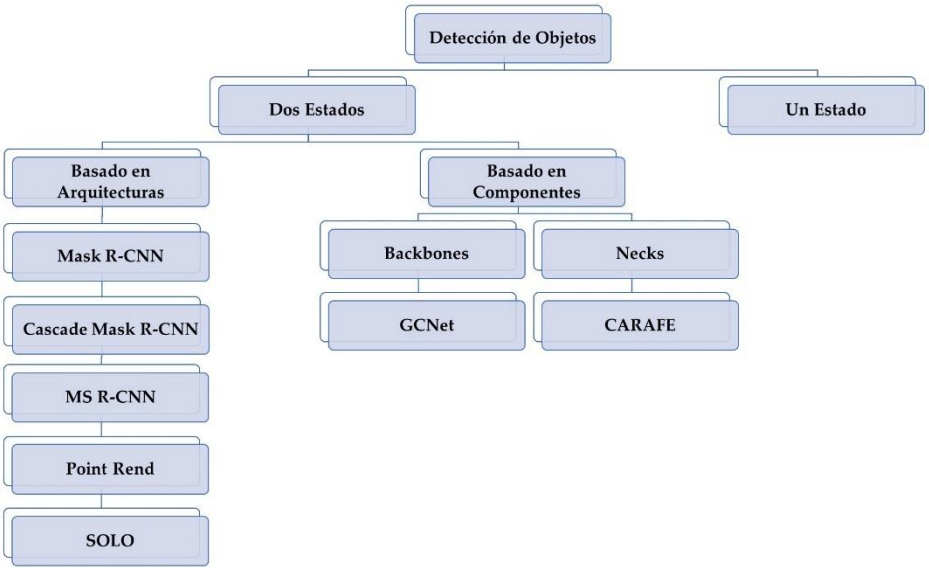


Figura 2-8: Modelos de detección de objetos de dos estados utilizados en esta investigación.

2.2.4.1. Fundamentos en los modelos de detección de objetos con DL.

El objetivo de esta sección es comprender los elementos críticos de cómo funcionan estos modelos, por qué son relativamente lentos pero poderosos y el proceso de compartir características mejora el detector de dos etapas.

- R-CNN [39]: se presentó el primer sistema exitoso para la localización, clasificación y segmentación de objetos, tomando de las imágenes originales alrededor de 2000 parches, conocidos como *regions*. Luego calcula la característica para cada propuesta usando una CNN y

finalmente clasifica cada región usando máquinas de vectores de soporte lineales específicas de cada clase.

- Fast R-CNN [40]: R-CNN es lento porque cada región propuesta pasa por una CNN sin compartir cálculos. En Fast R-CNN se pasa toda la imagen por una CNN. Se introduce el ROI *pooling* como una concatenación de entrada a salida de las características extraídas de cada región propuesta y se alimenta a una capa completamente conectada durante la predicción de categorías, con dos salidas: una probabilidad *softmax* y un desplazamiento de regresión de caja delimitadora por clase.
- Faster R-CNN [41]: Los modelos anteriores de detección de objetos todavía tienen un cuello de botella con la búsqueda selectiva, que tiene una mecánica letárgica y procesos que consumen mucho tiempo y que afectan el rendimiento de la red. En Faster R-CNN se propuso el concepto de redes de propuesta de regiones (RPN): se colocan sobre las características de la CNN, luego se reformulan utilizando ROI y se clasifican, y se llevan a cabo ambas tareas de caja delimitadora.
- Mask R-CNN [42]: se introdujo para predecir máscaras de segmentación en cada ROI con una *Fully Connected Network* pequeña en cada una. Este modelo extiende Faster R-CNN, agregando una nueva rama paralela a la rama existente para clasificar y usar la caja delimitadora. Este modelo aumenta ligeramente el costo computacional, pero sigue siendo un sistema rápido y permite una experimentación rápida. Mask R-CNN agrega una nueva rama a la arquitectura de Faster R-CNN para predecir máscaras de segmentación.

2.2.4.2. Componentes comunes en las arquitecturas de modelos de detección de objetos.

Los componentes estándares en arquitecturas de detección de objetos de dos etapas son:

- *Backbone*: La red toma una imagen como entrada y extrae el mapa de características sin la última capa completamente conectada. El *backbone*

puede ser una red neuronal pre-entrenada.

- *Neck*: Después del *backbone*, la capa *neck* extrae mapas de características más elaborados de diferentes etapas.
- *DenseHead*: Este componente funciona en ubicaciones densas de mapas de características. Un ejemplo es RPN, donde las cajas de anclaje se generan a partir de puntos de anclaje que se encuentran en los mapas de características. Las escalas y las relaciones de aspecto son elementos cruciales que se utilizan para crear cajas candidatas.
- *ROIExtractor*: Este componente extrae características de ROI utilizando técnicas de ROI *Pooling* y ROI *Aling*, lo que permite transformar celdas de destino no uniformes al mismo tamaño.
- *ROIHead*: Este componente toma características de ROI en una tarea específica, como la clasificación/regresión de cajas delimitadoras y la predicción de máscaras en la segmentación de instancias.

2.3. Herramientas y materiales

2.3.1. Python

Python es un lenguaje de programación de alto nivel potente y fácil de aprender, con una sintaxis simple y concisa que puede utilizarse en una amplia gama de tareas, que van desde el desarrollo web hasta el aprendizaje automático (ML). Tiene una biblioteca estándar que incluye una amplia gama de estructuras de datos eficientes, además de un sistema de programación orientado a objetos que permite a los desarrolladores crear código reutilizable y extensible [43].

Python es un lenguaje interpretado, lo que significa que no necesita ser compilado antes de su ejecución. Esto hace que sea ideal para scripting y desarrollo rápido de aplicaciones.

Su creador fue Guido Van Rossum en la década de los 90 y actualmente es administrado por *Python Software Foundation* bajo una licencia de código abierto,

llamada *Python Software Foundation License* [44].

Su selección para esta investigación se debió a su amplia gama de facilidades para el ML, la ciencia de datos y la visión por computadoras. Con este lenguaje se puede corregir y eliminar datos incorrectos, extraer y seleccionar características relevantes de los mismos y visualizarlos mediante el uso de tablas y gráficas. Ingenieros e investigadores de diferentes campos de la ciencia también utilizan clasificadores basados en Python para tareas como clasificación imágenes, texto, reconocimiento del habla, reconocimiento facial y DL.

Se utilizó Python en su versión 3.10.11.

2.3.2. Pytorch

Pytorch es una biblioteca de ML de código abierto que permite a los desarrolladores crear modelos de ML, especialmente redes neuronales, de forma rápida y eficiente.

Pytorch está basado en la biblioteca Torch, que fue desarrollada originalmente por el laboratorio de investigación de IA de Facebook. Esta biblioteca es compatible con Python y permite la ejecución de cálculos en GPU para acelerar el rendimiento.

Pytorch se utiliza para una amplia gama de aplicaciones de ML, incluyendo la visión por computadora, el procesamiento del lenguaje natural y el ML de refuerzo [45].

En esta investigación se utilizó Pytorch en su versión 2.0.1.

2.3.3. Anaconda

Anaconda es una distribución de Python y R que incluye paquetes científicos, bibliotecas y herramientas para la ciencia de datos. Es la herramienta elegida por los científicos de datos de todo el mundo.

Anaconda es popular por su escalabilidad, seguridad y simplicidad. Es fácil de instalar y usar, y proporciona una amplia gama de paquetes y bibliotecas para la ciencia de datos [46].

Permite desarrollar e implementar modelos de visión por computadoras, ML y DL rápidamente proporcionando herramientas para:

- La recopilación y análisis de datos.
- Gestión de entornos de paquetes con Conda, lo que facilita el control de diferentes versiones de software y paquetes para diferentes proyectos.
- Compatibilidad con múltiples plataformas, documentación y soporte amplio y una licencia gratuita.

2.3.4. Spyder como IDE

Spyder es un entorno de desarrollo integrado (IDE) de código abierto para Python, diseñado para científicos, ingenieros y analistas de datos. Es una herramienta poderosa que proporciona una combinación única de funcionalidad de edición, análisis, depuración y análisis de rendimiento de un entorno de desarrollo completo con las capacidades de análisis de datos de Python. Sus características principales son:

- Edición de código: Spyder proporciona un editor de código potente y extensible que incluye funciones como resaltado de sintaxis, autocompletado, finalización de código y navegación de código.
- Análisis de datos: Spyder incluye una amplia gama de herramientas para el análisis de datos, como un explorador de datos, un visor de gráficos y una calculadora numérica.
- Depuración: Spyder proporciona un potente depurador que permite a los desarrolladores rastrear el flujo de ejecución de su código y encontrar errores.
- Análisis de rendimiento: Spyder incluye herramientas para analizar el rendimiento de su código, lo que le ayuda a identificar áreas que pueden optimizarse.

Además de estas características principales, Spyder también incluye una serie de otras funciones que lo convierten en una herramienta versátil para el desarrollo de Python. Estas funciones incluyen integración con Anaconda, soporte para

múltiples lenguajes y ajustabilidad [47].

2.3.5. MMDetection

Por lo general las tareas de detección son más complejas que las de clasificación y llevan a diferentes implementaciones con diferentes resultados. Para reportar un resultado consistente se adoptó la herramienta de código base MMDetection, la cual provee implementaciones integradas para la detección de objetos y la segmentación de instancias basadas en Pytorch.

Esta herramienta pertenece al proyecto MMLab, un proyecto de código abierto para investigadores académicos y de la industria. Las características fundamentales de MMDetection son su diseño modular, soporte para múltiples entornos de trabajo de detección, una alta eficiencia que permite que todas sus operaciones básicas se ejecuten en la GPU y que las velocidades de entrenamiento sean altas, así como una constante actualización con las principales y nuevas investigaciones del estado del arte [48].

La versión utilizada para MMDetection fue 2.27 y de MMCV 1.6.

2.3.6. Colab

Google Colab se utilizó en esta investigación, el cual es un producto de *Google Research* que le permite a cualquier persona escribir y ejecutar código de Python a través del navegador y es muy adecuado para el ML y el análisis de datos. Los recursos asignados fueron una plataforma con un sistema operativa Linux y una GPU NVIDIA A100-SXM4 con 40 gigabytes de memoria RAM.

2.3.7. Hardware utilizado

El equipo utilizado durante todas las fases de la investigación fue una PC con un procesador Intel(R) Core(TM) i5-8400, con una velocidad de CPU a 2.80 GHz, con 16 GB de RAM y una tarjeta de video NVIDIA GeForce GTX 1070, con 8 GB de RAM dedicada al video.

2.3.8. Fuente de datos

Las bases de datos de imágenes utilizadas en esta investigación fueron las

siguientes:

a) REFUGE [49]: El *Retinal Fundus Glaucoma Challenge* fue el primer desafío sobre la evaluación del glaucoma a partir de fotografías del fondo de retina y es uno de los conjuntos de datos públicos más extensos disponibles para la segmentación de copa/disco. Consta de 1200 imágenes retinianas en formato JPEG. Se utilizaron dos dispositivos: una cámara de fondo de ojo Zeiss Visucam 500 con una resolución de 2124 x 2056 píxeles (400 imágenes) y una Canon CR-2 con una resolución de 1634 x 1634 píxeles (800 imágenes). La mácula y el disco óptico son visibles en cada imagen, centrados en el polo posterior.

b) G1020 [50]: Se recopiló un nuevo conjunto de datos públicos para la segmentación de imágenes de la copa/disco en una clínica privada en Kaiserslautern, Alemania, entre los años 2005 y 2017. Las imágenes tienen un campo de visión de 45 grados después de las gotas de dilatación. Los expertos marcaron los límites del disco óptico y de la copa y las anotaciones de las cajas delimitadoras utilizando Labelme [51], una herramienta de anotaciones gratuita y de código abierto. Las imágenes se almacenan en formato JPG con tamaños entre 1944 x 2108 y 2426 x 3007 píxeles.

c) ORIGA [52]: Conjunto de datos utilizado para el análisis del glaucoma. Este conjunto de datos se utilizó para segmentar la atrofia peripapilar en su clasificación, alfa y beta. Se seleccionaron 267 imágenes de fondo de ojo en formato JPG, con un tamaño de 3072 x 2048. El proceso de anotación se realizó a través de Roboflow [53], una plataforma de código abierto para manejar imágenes con técnicas de visión artificial, bajo la supervisión de especialistas del Instituto Mexicano de Oftalmología.

d) DDR [54]: Este es el conjunto de datos principal para esta investigación, ya que tiene la lesión más compleja de detectar. Consta de 13673 imágenes de fondo de ojo en color, de hospitales de China entre los años 2016 y 2018. Estas imágenes fueron tomadas con una amplia gama de cámaras de fondo de ojo, 42 en total con 45º de campo de visión, y en formato JPG. De ese número de imágenes, se seleccionaron 757 para anotar las lesiones de RD,

MA, EX, SE y HE.

2.4. Estado del arte

La RD afecta a los pequeños vasos sanguíneos de la retina. Puede causar inflamación en la mácula, la parte central de la retina, o crecimiento anormal de nuevos vasos sanguíneos que pueden sangrar o tirar de la retina [23]. Dependiendo de la gravedad, aparecen diferentes biomarcadores de daño, como HE en forma de llama, MA, EX y manchas de algodón (SE) [22].

Por otro lado, en la examinación de imágenes de fondo de ojo, la región del disco óptico es comúnmente analizada para evaluar el glaucoma. Importantes marcadores que son extraídos de esta área son la RCD, el área del anillo neurorretiniano, así como ISNT, donde la región inferior debe ser mayor a la superior, esta a su vez que la nasal y esta que la temporal para completar la regla [22].

Estos marcadores no son los únicos mecanismos que permiten evaluar el glaucoma. Los estudios longitudinales sugieren que una gran zona beta de APP predice la progresión del glaucoma en pacientes con glaucoma crónico de ángulo abierto [55].

Las características descritas anteriormente son características morfológicas que se utilizan como parte de la evaluación del glaucoma y deben detectarse [56], y características asociadas a la retinopatía diabética que deben detectarse y localizarse. Tradicionalmente, esta tarea ha dependido de la inspección manual por parte de oftalmólogos cualificados, lo que requiere mucho tiempo, es subjetivo y propenso a errores humanos.

Por esta razón, una cantidad considerable de investigaciones propone la detección y segmentación del disco óptico para extraer información de él.

Se puede encontrar una amplia revisión de la literatura en [57], con métodos como los basados en conjuntos de niveles, en umbrales, en clústeres y en RNA, este último con una precisión más general y menos tiempo de procesamiento, pero

con un alto coste computacional. Un documento revisado recientemente [58] resume los conjuntos de datos públicos existentes, así como los métodos de ML y DL para el disco óptico (DO) y la copa óptica (CO). Se presentaron un total de 29 arquitecturas, y la mayoría de los trabajos emplean CNN, enfoques basados en U-Net y redes generativas adversariales (GAN), y solo uno utilizó una R-CNN, [59], donde los autores utilizaron una R-CNN más rápida y transformaron el cuadro delimitador predicho en una elipse vertical y no rotada. Ninguno de los modelos revisados utilizaba R-CNN para la tarea de segmentación, excepto uno.

Otros ejemplos de arquitecturas basadas en DL han sido liberadas. Ejemplos de arquitecturas U-Net revisadas pueden consultarse en [60]–[63], y segmentaciones basadas en GAN pueden verse en [64]–[67]. Otro trabajo reciente basado en CNN es el presentado por [68]. Los autores emplean una capa de convolución separable en profundidad y una entrada de pirámide de imágenes, con una puntuación de dados de 0,96 para DO y 0,89 para CO en el conjunto de datos REFUGE. Z. Tian et.al. en [69], propusieron una red convolucional gráfica que toma el mapa de características concatenado con los nodos gráficos como entrada para la segmentación de DO y CO, consiguiendo un índice Dice de 0.9776 y 0.9558, respectivamente. Y. Zheng et. al. en [70], presentaron una CNN multiescala para generar parámetros iniciales de contorno y evolución, reportando sobre el conjunto de datos REFUGE una intersección sobre unión (IoU) de 0.9669 para DO y 0.9361 para CO. Finalmente, J. Zhang et. al. emplearon un enfoque novedoso en imágenes láser de escaneo multicolor en [71], donde los autores producen anotaciones funcionales a través de un *crowdsourcing* no experto, que aprovecha un par de redes de regularización y segmentación.

Recientes avances en la visión por computadoras y el ML han introducido los algoritmos de detección de objetos, los cuales automáticamente identifican y localizan objetos dentro de una imagen. La detección de objetos tiene como objetivo contar objetos en una escena y determinar y rastrear sus ubicaciones exactas mientras los etiqueta con precisión. Estos modelos tienen como ventajas que son más resistentes a las oclusiones, los retos en la iluminación y las escenas

complejas.

La detección de objetos se ha utilizado ampliamente en imágenes médicas, específicamente en el sistema digestivo, respiratorio, cardiovascular, ocular y mamario [72]. La detección de objetos médicos es la tarea de identificar objetos médicos dentro de una imagen, y algunos ejemplos son la detección y localización de aneurismas intracraneales [73], la detección de fracturas en radiografías de muñeca [74], relacionadas con el sistema digestivo [75]–[78], y en patología [79], [80]. Sin embargo, muchos de estos trabajos son de detección de una sola clase; ejemplos de detección de múltiples clases, el objetivo de esta investigación, son la detección de células, la detección de lesiones en imágenes de tomografía computarizada y la gravedad del acné [81]–[85].

Algunos trabajos aplican técnicas de detección de objetos como en [86], para la localización de DO, donde utilizan un Faster R-CNN para lograr una precisión satisfactoria y una alta velocidad en comparación con otros trabajos, y el de [87] detecta el glaucoma a través de un Faster R-CNN. Los autores [88] propusieron una Red de Enfoque de Regiones, donde se diseñó una nueva rama de máscara multiclase. En el trabajo [89] mostraron un estudio comparativo detallado entre cuatro algoritmos de DL diferentes: YOLOv2, YOLOv2 reducido, YOLOv3 y Mask RCNN, donde en términos de precisión, YOLOv3 muestra el mejor rendimiento; sin embargo, en términos de IoU, Mask R-CNN muestra los mejores resultados. Los autores, en [90], propusieron un Mask R-CNN basado en Densenet-77 para abordar las imágenes retinianas borrosas, reportando una IoU de 0,972. Un problema de desplazamiento de dominio esencial entre diferentes conjuntos de datos fue propuesto por Y. Guo et. al. [91], a través de un Faster R-CNN adaptativo de grueso a fino para la segmentación conjunta de DO y CO. En el trabajo de [92], utilizaron un Mask R-CNN simple de dos etapas, que primero detecta y corta alrededor del disco óptico, luego introduce la imagen recortada con la original en la nueva red de detección utilizando diferentes escalas. Por último, [93], en lugar de detectar una caja delimitadora, estiman directamente los parámetros de una elipse que es suficiente para capturar la morfología de cada región de DO y CO para calcular la

RCD.

Se ha investigado menos sobre la APP, pero la evolución ha progresado desde técnicas de visión por computadora como [94], [95], hasta técnicas de DL como las CNN y la arquitectura U-Net [96] en los siguientes trabajos [97], [98]. Chai et al. propusieron un enfoque interesante que combina múltiples características relacionadas con el glaucoma (DO, CO, APP, capa de fibras nerviosas de la retina) y el conocimiento del dominio a través de una red neuronal multitarea [99].

En la RD, muchas investigaciones se han centrado en el diagnóstico de la enfermedad a nivel de imagen, la clasificación del estado y la localización de las características relacionadas. Los siguientes trabajos se centraron en la extracción de lesiones de la retina asociadas a la RD, concretamente MA [100], MA y HE en [101], EX y SE en [102], y todas estas lesiones juntas en [103]. Los siguientes estudios son investigaciones más completas, ya que clasifican y localizan las lesiones [104]–[106].

Todos estos estudios previos se centran en una sola enfermedad. Es bien sabido que la retina es el único lugar de nuestro cuerpo donde podemos ver patologías en tiempo real que afectan a la retina [107], y son varias. Por ello, nuevas investigaciones están comenzando a clasificar, localizar y calificar diferentes afecciones dentro de la retina simultáneamente.

Son et al. en [108] desarrollaron un sistema de DL para el cribado de doce anomalías retinianas. Utilizan doce redes neuronales profundas, una para cada enfermedad. Wang et al. en [109] y Karthikeyan et al. en [110] crearon un sistema para la clasificación multietiqueta, que es una clasificación a nivel de imagen, y se proporcionan algunos indicios de cómo se realizó la predicción. Un enfoque similar sigue en [111].

El progreso va más allá con el trabajo de [112], que fueron capaces de detectar 39 enfermedades con imágenes de fondo de ojo utilizando cuatro grupos de CNN y un modelo de detección de objetos, y [113] predijeron 47 biomarcadores sistémicos como variable de resultado a partir de fotografías retinianas, incluyendo

el sexo, la edad, la presión arterial, los perfiles lipídicos, la cantidad y otros. Emplearon un total de 47 algoritmos de DL.

2.5. Propuesta de solución

En las extensas revisiones del estado del arte se utilizan muchas técnicas, principalmente las asociadas a las CNN, las arquitecturas U-Net y los modelos basados en GAN. Se pueden observar dos ramas principales: una relacionada con la extracción de características, generalmente con un modelo para una enfermedad específica, y la otra centrada en la clasificación multietiqueta, que deriva en la identificación de múltiples enfermedades, pero que requiere varios modelos de DL.

Estos modelos se entrenan y prueban en diversos conjuntos de datos que deben ser aumentados. Por otro lado, la detección de objetos no se ha abordado ampliamente en esta área, por lo que la motivación de esta investigación es en primer lugar realizar un análisis detallado del comportamiento de diferentes modelos de detección de objetos en la detección y segmentación del DO y CO, enriqueciendo la literatura actual al tiempo que se establecen las métricas típicas de estos modelos como base para futuras comparaciones, y como segunda etapa, haciendo uso de uno de estos modelos, realizar la detección simultánea de múltiples lesiones relacionadas con el glaucoma y la RD en imágenes de retina.

Otra razón es que estos nuevos modelos se prueban tradicionalmente en conjuntos de datos clásicos como *Common Objects in Context* (COCO) [114] y *The Pascal Visual Object Classes* [115], y se emplearán en imágenes de la retina del fondo del ojo, abordando el número de imágenes necesarias para entrenar redes neuronales profundas, probando modelos en conjuntos de datos reducidos y completos y realizando comparaciones estables.

Adicionalmente, en lugar de dar un diagnóstico, proporcionar características de cada enfermedad que pueden ser útiles para los no especialistas, especialmente en regiones remotas, y por lo tanto puedan remitir a un especialista que pueda complementar el diagnóstico con otros factores de riesgo e información sistémica. Para lograrlo, es crucial manejar objetos pequeños y clases desequilibradas; un

problema recurrente en la detección de objetos que aparece cuando una clase es mucho más grande que otra y cuando los objetos son más pequeños de 32 x 32 píxeles. La Tabla 2-1 muestra la declaración de significancia de esta investigación.

Tabla 2-1: Declaración de significancia del presente trabajo de investigación.

Problema	• La interpretación de la segmentación del DO y la CO es crucial para el diagnóstico del glaucoma. Unos resultados precisos pueden marcar la diferencia entre una buena y una mala predicción. • Los médicos pueden detectar muchas enfermedades a la vez en la retina, pero los sistemas de asistencia de IA suelen detectar una sola enfermedad.
¿Qué es conocido?	• Las redes neuronales profundas se utilizan para la segmentación, centrándose en los modelos codificador-decodificador. • Gran trabajo de preprocesamiento y posprocesamiento. • Los flujos de trabajo se basan en la extracción previa de la región de interés para realizar la segmentación en esa área recortada. • La mayoría de los trabajos se centra en extraer lesiones de una sola enfermedad. • Predicción multietiqueta y clasificación multiclase utilizando más de una red neuronal profunda para realizar la tarea.
¿Qué aporta esta investigación?	• Evaluar el estado del arte de los nuevos modelos de detección de objetos con un enfoque de dos etapas, destacando la mejor precisión media. Estos modelos unifican la tarea de detección y segmentación. • Abordar la pregunta tradicional de cuántas imágenes son necesarias para entrenar un modelo de red neuronal profunda. Experimentar con el rendimiento en un subconjunto y en el conjunto de datos completo. • El efecto de la técnica de aumento de datos a múltiples escalas y la importancia de una configuración correcta de la escala de las anclas para la localización de objetos. • Implementar una estrategia para evitar los falsos negativos con datos parcialmente etiquetados. • Justificación de una arquitectura específica para manejar el desequilibrio en la distribución. • Abordar los desafíos de la detección de objetos pequeños implementando la distancia Wasserstein normalizada. • Mejorar el posprocesamiento con una selección refinada de cajas delimitadoras. • Establecer una comparación de vanguardia sobre un conjunto de datos específico.

3. Metodología

De manera general se pretende seguir un flujo de trabajo en el cual se integren fases como parte del diseño de un entorno de trabajo que agrupe la detección de lesiones en imágenes de retina asociadas a enfermedades oculares (RD y Glaucoma).

Para darle cumplimiento a los objetivos de la investigación se diseñaron cuatro estrategias que nos guiaron hacia su conclusión:

- Estrategia 1: Identificación, desarrollo y prueba de algoritmos de DL para detectar lesiones, validando su desempeño y estableciendo una comparación entre ellos.
- Estrategia 2: Diseño de un marco de trabajo que permita, la ingestión de datos con su correspondiente preprocesamiento y transformación; así como la integración de modelos de DL en un único ecosistema, que permita la detección de múltiples lesiones bajo un solo entorno.
- Estrategia 3: Desarrollo de la integración entre herramienta de software y marco de trabajo para la visualización de resultados y uso en instituciones de salud.
- Estrategia 4: Determinación del grado de aproximación del marco de trabajo creado con los resultados del estado del arte.

3.1. Segmentación de instancias

Para darle comienzo a la primera etapa se seleccionaron varios modelos de segmentación de instancias para extraer tanto el DO como la CO. La tarea de extracción está relacionada con la segmentación de instancias, que permite detectar y localizar un objeto en una imagen. El objetivo de la segmentación de instancias es obtener objetos de la misma clase divididos en diferentes instancias; aunque por concepto, el disco y la copa son clases diferentes, tienen una forma muy similar y se superponen. Por esta razón, es necesario extraerlos por separado. Los modelos de detección de objetos pueden abordar este problema y se cubrirán aquí, evaluando el rendimiento de las arquitecturas recientes.

En la Figura 3-1 se puede ver la metodología seguida en esta etapa, la cual nos permite analizar el comportamiento de los modelos de detección de objetos basados en diferentes enfoques de arquitecturas, diferentes enfoques en el *Backbone* y en el *Neck*. El flujo de trabajo general comienza con la recuperación de las imágenes y su anotación correspondiente. A continuación, se configura la experimentación con y sin datos aumentados a múltiples escalas y con una configuración adecuada de las anclas antes de entrenar y predecir el área segmentada.

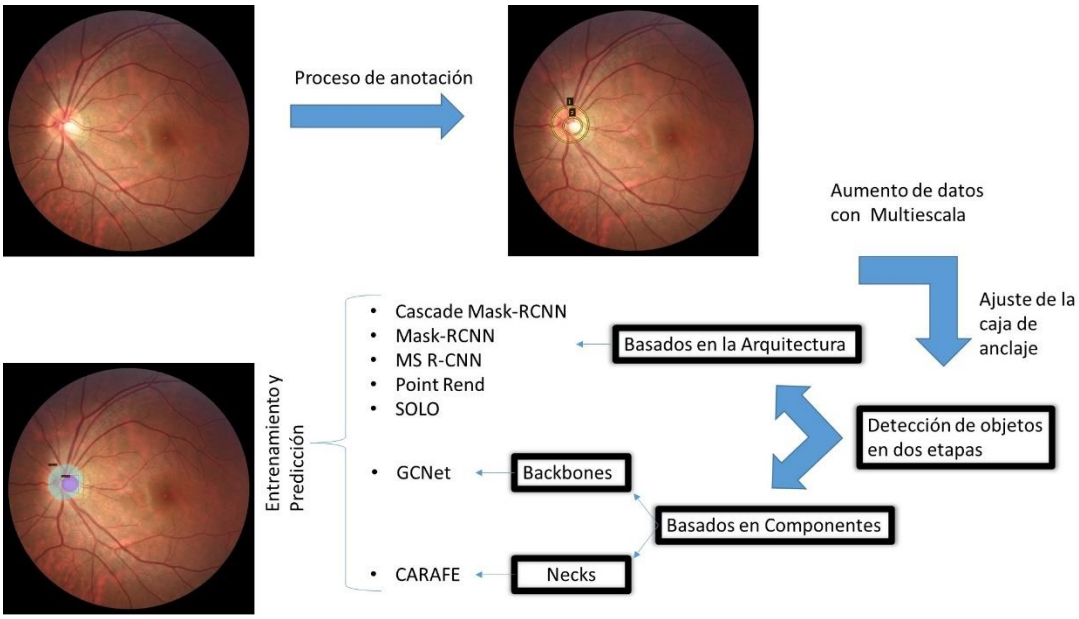


Figura 3-1: Flujo de trabajo propuesto para la segmentación del DO y CO con diferentes modelos de detección de objetos.

3.1.1. Modelos

3.1.1.1. Cascade R-CNN

Cascade R-CNN es una arquitectura de detección de objetos de varias etapas en la que se utilizan umbrales de IoU crecientes en un detector de orden secuencial, utilizando la salida de uno como entrenador del siguiente, mejorando la calidad, garantizando un conjunto de entrenamiento positivo y minimizando el sobreajuste [116]. Esta arquitectura es una extensión de Faster R-CNN, y la obtención de máscaras se puede abordar de dos maneras: colocando la rama de segmentación al principio o al final de Cascade R-CNN o en cada etapa. Esta última maximiza la

diversidad de muestras utilizadas para aprender la tarea de predicción de máscaras.

3.1.1.2. Mask Scoring R-CNN

Para la mayoría de los modelos de segmentación de instancias, la confianza de la clasificación de instancias se utiliza como una puntuación de calidad de la máscara. Sin embargo, en la práctica, la máscara de instancias y el *ground truth* no suelen estar bien correlacionadas con las puntuaciones de clasificación. La idea detrás de esta arquitectura es tomar la característica de la instancia y la máscara predicha correspondiente juntas para regresionar la IoU de la máscara a través de un cabezal MaskIoU [117]. En esta propuesta, la máscara predicha y la característica ROI se toman como entrada para el cabezal MaskIoU.

3.1.1.3. PointRend: Segmentación de imágenes como renderizado

Este módulo nos aporta flexibilidad para realizar predicciones de segmentación basadas en puntos en ubicaciones seleccionadas de forma adaptativa mediante un algoritmo de subdivisión iterativo. El modelo PointRend proporciona bordes de objetos nítidos en regiones que han sido suavizadas en exceso por métodos anteriores [118]. PointRend selecciona un conjunto de puntos para realizar la tarea y predice cada punto individualmente con un pequeño perceptrón multicapa, utilizando características interpoladas calculadas en estos puntos. Este proceso se aplica secuencialmente para optimizar las regiones conflictivas de la máscara predicha.

3.1.1.4. CARAFE

Content-Aware ReAssembly of Features (CARAFE), explota un gran campo de visión, agregando información contextual. Permite el manejo de contenido específico de la instancia, generando *kernels* adaptables instantáneamente, y es liviano y rápido de calcular [119]. CARAFE está compuesto por dos componentes principales: el módulo de predicción del *kernel* genera *kernels* de reensamblaje de forma consciente del contenido. Por el contrario, el módulo de reensamblaje consciente del contenido reconstruye las características de cada *kernel* de reensamblaje dentro de una región local con una función específica.

3.1.1.5. GCNet

Global Context Network (GCNet) fue propuesto para mejorar NLNet, cuya tarea es capturar dependencias de largo alcance a través de la agregación de contexto global específico de la consulta a cada posición de la consulta. La mejora se estableció en tres pasos: un marco general que obtiene una mejor instanciación basada en una formulación independiente de la consulta [120].

3.1.1.6. SOLO

Segmentando objetos por ubicación [121] introduce las "categorías de instancia", un enfoque que asigna categorías a cada píxel dentro de una instancia según su ubicación y tamaño. Este enfoque transforma la segmentación de instancias en un problema de clasificación de un solo paso. El modelo propuesto divide la imagen de entrada en una cuadrícula uniforme y, si el centro de un objeto cae en una celda de la cuadrícula, esta predice la categoría semántica y segmenta esa instancia de objeto.

3.1.2. Anotación y preprocesamiento

El preprocesamiento de datos en el DL es el proceso de preparar los datos para que los puedan utilizar los modelos de DL. Esto incluye tareas como la limpieza, la normalización y la transformación de los datos.

El preprocesamiento de datos es una parte esencial del DL, ya que puede tener un impacto significativo en el rendimiento de los modelos. Los datos limpios y bien preparados pueden ayudar a los modelos a aprender más rápido y con mayor precisión.

El marco de trabajo con el que se trabajó fue MMDetection, el cual provee los modelos antes mencionados. Este marco de trabajo soporta el conjunto de datos de estilo COCO y es un conjunto de datos de detección de objetos, segmentación y subtitulado a gran escala. Es crucial para la localización precisa de los discos ópticos y de copa. Se utilizó el software VGG Image Annotator (VIA) [122] para generar las anotaciones de la ROI para un procedimiento de entrenamiento adecuado y correcto. Este software es de código abierto y es una plataforma de

anotación manual independiente y directa para imágenes, audio y vídeo. El conjunto de datos REFUGE se configuró a través de estas herramientas y luego se exportó en formato COCO.

Un ejemplo puede verse en la Figura 3-2. Se seleccionó una forma elíptica ya que es la que mejor se ajusta a los discos ópticos y de copa. Se utilizó el *ground truth* original de ambos conjuntos de datos como guía.

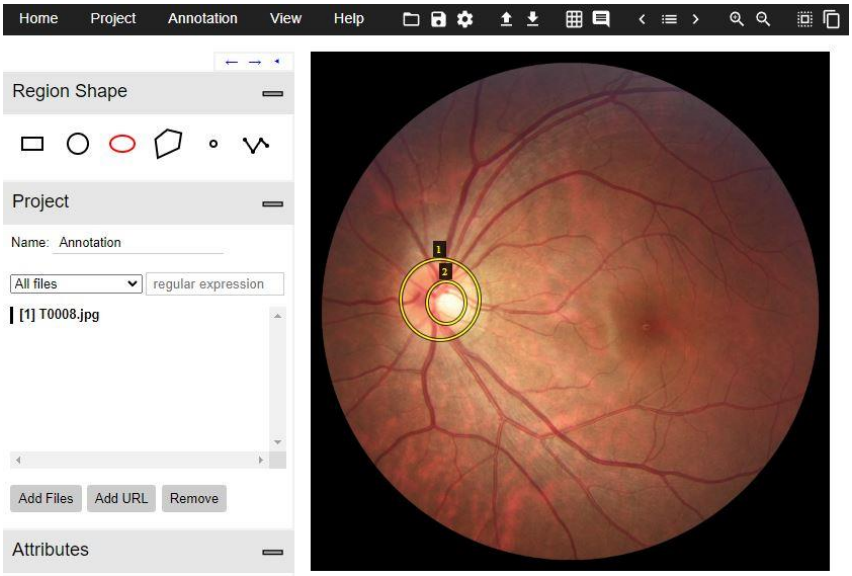


Figura 3-2: Ejemplo de anotación de una imagen REFUGE. Se seleccionó la forma elíptica. El número uno anota la clase de disco y el número dos anota la clase de copa.

El preprocesamiento y la ampliación de imágenes son siempre una parte vital del comportamiento exitoso de una red neuronal; sin embargo, las transformaciones agresivas no siempre conducen a mejores resultados. En esta fase de la investigación se realizaron algunos pasos. Primero, se redimensionaron las imágenes, adoptando un esquema simple de aumento de datos basado en el entrenamiento a múltiples escalas, en el que las imágenes se tomaron en tamaños entre 1333 x 640 y 1333 x 960 con un paso de 32 entre cada una de ellas. Este enfoque muestra un alto rendimiento en términos de precisión media (AP), caja delimitadora y máscaras con respecto a un tamaño fijo. Luego, se realizó un giro aleatorio seguido de una normalización basada en la media y la desviación estándar de ImageNet [36], comúnmente utilizado como aprendizaje de transferencia para acelerar el proceso de entrenamiento.

3.2. Detección múltiple de lesiones

Esta etapa de la investigación está asociada a la estrategia dos y se enfrenta a diferentes problemas en múltiples niveles que deben abordarse. Busca la creación de un ecosistema que permita la detección de lesiones en la retina que pertenecen a dos enfermedades diferentes, RD y Glaucoma, bajo un único modelo de detección de objetos.

La metodología propuesta tiene dos partes principales, un proceso de anotación y la detección de lesiones a múltiples escalas y tamaños, con mejoras claves en cada etapa dentro del flujo del proceso. Se eligió un modelo de detección de objetos de dos fases para alcanzar el objetivo, ya que puede disminuir una gran cantidad de ejemplos negativos en el proceso de extracción de características [41].

La Figura 3-3 describe el proceso general, comenzando con la anotación del conjunto de datos, donde se utilizaron técnicas de etiquetado suave y visión por computadora. Se utiliza un modelo Cascade R-CNN [116]. Para extraer características, se utiliza Resnet50 [123] como *Backbone* en combinación con *Feature Pyramid Network* (FPN) [124] en el módulo *Neck*. En RPN se adoptó una métrica de Distancia de Wasserstein Normalizada, inicialmente inspirada en [125]; mientras tanto, se empleó *Online Hard Example Mining* (OHEM) para muestrear ejemplos positivos/negativos durante el entrenamiento en regiones de interés. Alternativamente, se empleó una función de pérdida asimétrica, en reemplazo de la entropía cruzada, para mejorar la selección de muestras. Finalmente, se aplicó una técnica mejorada de supresión de no máximo (NMS) en el paso de posprocesamiento para mitigar resultados duplicados o superpuestos.

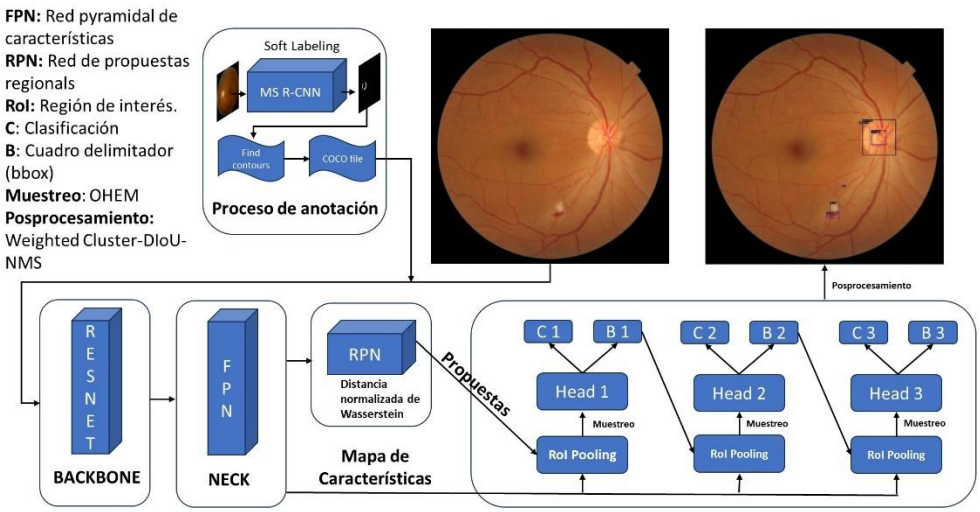


Figura 3-3: Diagrama de flujo de la investigación propuesta. Dos fases, iniciando por el etiquetado suave, seguido de la detección a través de mejoras dentro del modelo Cascade R-CNN.

3.2.1. Proceso de anotación

La anotación de imágenes es la práctica de etiquetar imágenes para entrenar modelos de IA y ML. A menudo implica que anotadores humanos utilicen una herramienta de anotación de imágenes para etiquetar imágenes o etiquetar información relevante, por ejemplo, asignando clases relevantes a diferentes entidades en una imagen. Los datos resultantes, también conocidos como datos estructurados, se envían a un algoritmo de ML, que a menudo se entiende como entrenamiento de un modelo.

Diferentes tareas requieren que los datos se anoten en diferentes formas. Las tareas complejas como la segmentación y la detección de objetos requieren que los datos tengan anotaciones de mapa de pixeles y anotaciones de cuadros delimitadores respectivamente.

Dado que esta investigación pretende detectar múltiples lesiones relacionadas con diferentes enfermedades, una decisión lógica podría ser fusionar conjuntos de datos con anotaciones de diferentes lesiones. Sin embargo, esto crea un problema de datos parcialmente etiquetados, que aparece cuando hay anotaciones faltantes en un conjunto de datos de detección etiquetado, lo que significa la presencia de un falso negativo en el *ground truth* (objetos presentes en el conjunto de datos, pero

no anotados). En consecuencia, dañará el proceso de aprendizaje de los modelos de detección de objetos porque todo lo que no se etiquetó o no coincidió con un ancla se considerará fondo.

El proceso por el cual un modelo etiqueta imágenes por sí solo a menudo se denomina etiquetado asistido por modelo o *soft labeling*, el cual fue empleado en el conjunto de datos DDR, con un modelo entrenado sobre otro conjunto de datos, completando las anotaciones con predicciones [126].

El modelo de detección de objetos Mask Scoring R-CNN, previamente analizado, se entrenó en los conjuntos de datos ORIGA y G1020 para la segmentación de la APP y copa/disco, respectivamente. Se seleccionó este modelo porque penaliza las puntuaciones de predicción cuando la clasificación es correcta, pero la máscara de segmentación no lo es.

Inicialmente, el conjunto de datos DDR proporcionaba una carpeta para cada máscara de segmentación relacionada con las lesiones (MA, EX, SE, HE). Después de la operación de *soft labeling*, se agregaron tres carpetas más con APP alfa, APP beta y máscaras para la copa y el disco; la Figura 3-4 es una muestra de una imagen y una máscara para cada lesión.

El próximo paso es la creación de un fichero de anotación con el estilo COCO,

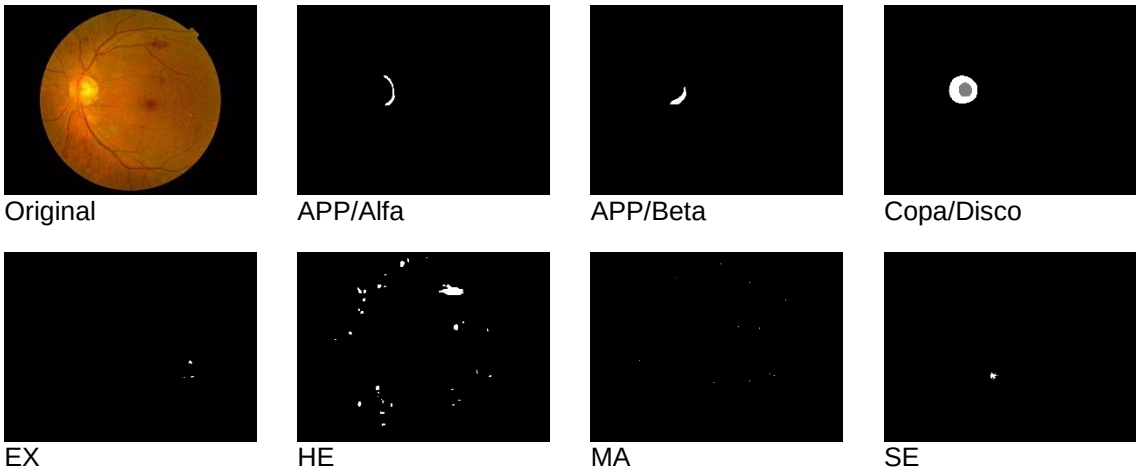


Figura 3-4: Ejemplo de imagen de entrenamiento y todas sus máscaras de segmentación. El modelo MS R-CNN generó APP/Alfa, APP/Beta y Copa/Disco como máscaras predichas. Ex, HE, MA y SE son máscaras originales en el conjunto de datos DDR.

para que pueda ser utilizado por el marco de trabajo de MMDetection.

El procedimiento comenzó iterando sobre cada carpeta de máscaras

seleccionando cada imagen y encontrando el contorno de cada máscara de lesión. OpenCV [127] proporciona una biblioteca de visión por computadora optimizada en tiempo real, y se utilizó la función findContours de esta biblioteca. Una descripción del algoritmo es el siguiente:

- La función findContours toma una máscara como entrada y devuelve una lista de contornos.
- El primer paso es encontrar los contornos en la máscara utilizando la función cv2.findContours. Esta función toma tres parámetros: la máscara, el modo de recuperación y el método de aproximación. El modo de recuperación especifica cómo se devuelven los contornos. El método de aproximación especifica cómo se simplifican los contornos. En este caso, estamos utilizando el modo de recuperación externo y el método de aproximación simple.
- El siguiente paso es cerrar los contornos. Esto se hace recorriendo la lista de contornos y agregando el primer punto al final de cada contorno si el contorno tiene más de tres puntos. Esto asegura que todos los contornos estén cerrados.
- El paso final es devolver la lista de contornos.

Después de aplicar el algoritmo findContour, todas las lesiones de diferentes carpetas se localizaron bajo anotaciones de cuadros delimitadores. La Figura 3-5 muestra una imagen de muestra con cuadros delimitadores de color por lesión.

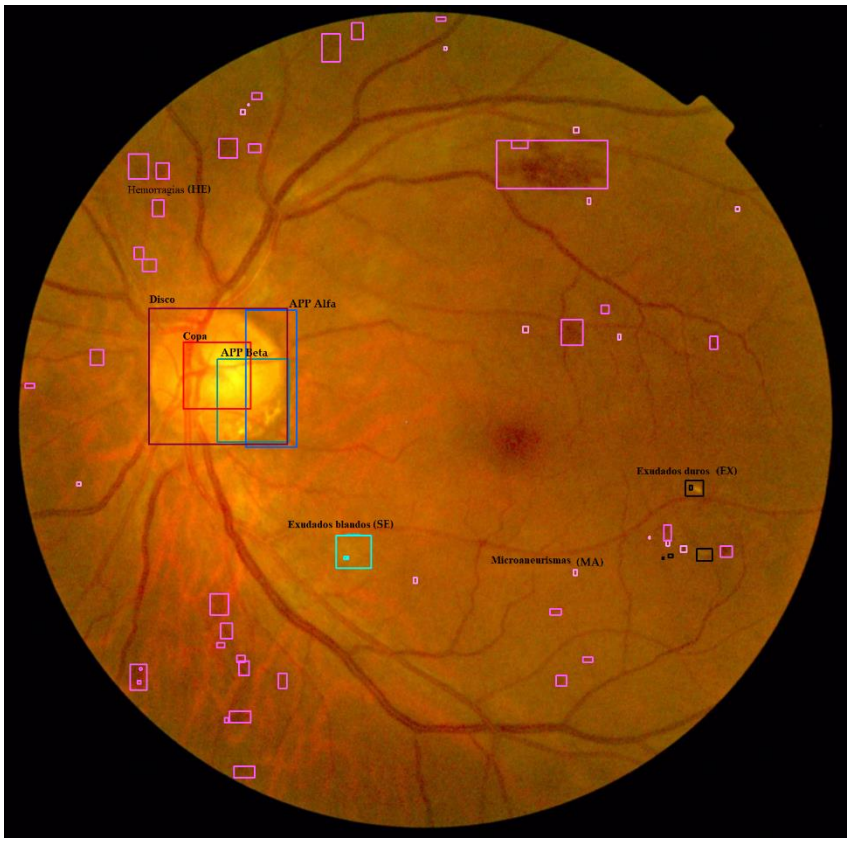


Figura 3-5: Imagen de ejemplo con cuadros delimitadores anotados por lesión. Se etiquetó una lesión por categoría para fines de aclaración. Se identificaron un total de ocho características. Relacionados con el glaucoma son Disco, Copa, APP Alfa y APP Beta. Relacionados con la retinopatía diabética son HE, MA, SE, EX.

3.2.2. Exploración de datos

La exploración de datos es clave para muchos procesos de ML y cuando hablamos específicamente de la detección y segmentación de objetos en imágenes, no existe una forma sencilla de realizar una exploración de datos sistémica. Para obtener una mejor comprensión de nuestros datos, evaluar la calidad de estos es un paso importante, sobre todo si se entrena con conjuntos de datos personalizados que son significativamente diferentes de los conjuntos de datos de referencia típicos como COCO.

Una vez establecido el conjunto de datos, se necesita un análisis profundo para comprender las características de la imagen y seleccionar el modelo y la configuración de hiperparámetros adecuados. Una de las tareas fue inspeccionar las relaciones de aspecto de las imágenes. La Figura 3-6 muestra la distribución.

Lo que se puede inferir de este gráfico es la presencia de una distribución trimodal, con la mayoría de las imágenes con relaciones de aspecto entre 1 y 1.5 y algunas cercanas a 2. La relación de aspecto es un hiperparámetro que debe ajustarse, y para cubrir el rango de valores identificados en la Figura 3-6, se tomaron los valores 0.5, 1.0, 1.5, 2.0 como vector para la variable anchor_ratios. La distribución es similar a la del conjunto de datos COCO, lo que permite un cambio de tamaño no destructivo seguido de un enfoque de relleno ligero.

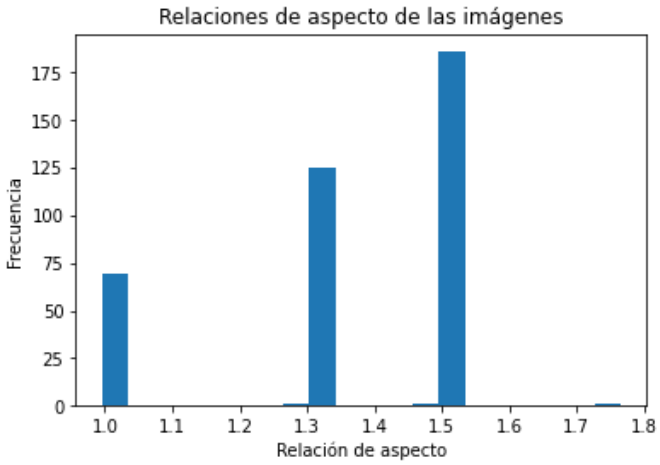


Figura 3-6: Relación de aspecto de los cuadros delimitadores en el conjunto de datos DDR.

A continuación, contar el número de objetos fue obligatorio para identificar las distribuciones de clases, véase la Figura 3-7. Como se puede observar, también existe un problema de desequilibrio en el área de los objetos.

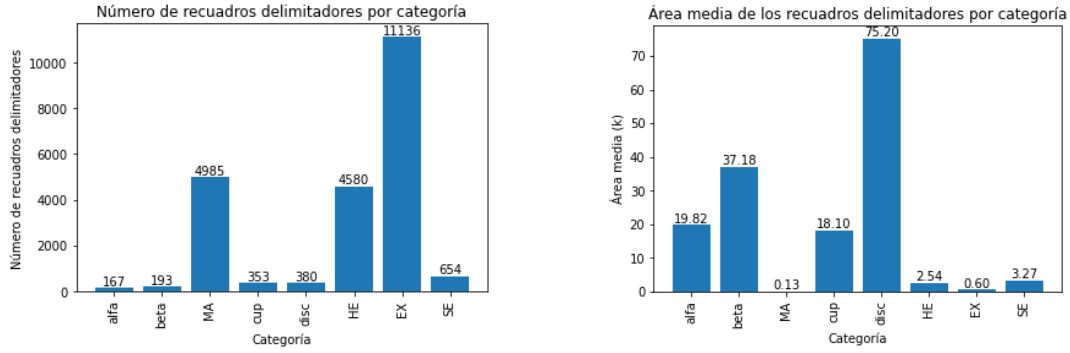


Figura 3-7: Conteo de cuadros delimitadores por clase (izquierda). Área media de cuadros delimitadores por clase (derecha).

Esta información es clave para la selección del modelo y, como se puede ver, hay presencia de objetos de diferentes tamaños y un problema de desequilibrio de

cuadros delimitadores. Eso fue una advertencia para evitar las técnicas de recorte porque se puede cortar en otros cuadros delimitadores o máscaras de segmentación más grandes, por lo que resolver un problema puede crear otro.

3.2.3. Preprocesamiento y aumento de datos

El preprocesamiento y el aumento de imágenes son cruciales para el rendimiento exitoso de una red neuronal. Sin embargo, las transformaciones agresivas solo a veces producen mejores resultados. En este estudio, se tomaron varios pasos para el entrenamiento. Después de cargar las imágenes y las anotaciones, se realizó un cambio de tamaño, utilizando un esquema de aumento simple a través de un entrenamiento a escala múltiple con tamaños que van desde 1333 x 640 hasta 1333 x 960. Este enfoque demuestra un alto rendimiento en el AP del cuadro delimitador en comparación con un tamaño fijo. Posteriormente, se aplica una operación de giro aleatorio, seguida de una normalización basada en la media y la desviación estándar de las imágenes en ImageNet [128], que se utiliza comúnmente para el aprendizaje por transferencia para acelerar el proceso de entrenamiento.

3.2.4. Marco de trabajo de detección de objetos

Las técnicas de detección de objetos pueden considerarse un marco de trabajo (*framework*), ya que abarcan múltiples redes neuronales profundas y metodologías que trabajan en conjunto para detectar y localizar objetos dentro de imágenes o videos. Esto enfatiza su naturaleza integral y sistémica. No se trata de un conjunto de herramientas aisladas, sino de una combinación sinérgica de elementos que convergen en un objetivo común: la identificación precisa y eficiente de objetos en imágenes y videos. Esto le permite a los investigadores y profesionales aprovechar los métodos y arquitecturas existentes o desarrollar nuevos para mejorar el rendimiento y cumplir con requisitos específicos.

3.2.4.1. Backbone

ResNet (Red Neuronal Residual) es una arquitectura de red neuronal convolucional profunda conocida por su aplicación exitosa en tareas de reconocimiento de imágenes [123]. Introduce bloques residuales, que permiten el

entrenamiento de redes profundas al mitigar el problema del desvanecimiento del gradiente. La arquitectura ResNet consta de varios bloques, cada uno de los cuales contiene múltiples bloques residuales. Estos bloques reducen progresivamente las dimensiones espaciales de la entrada al tiempo que aumentan el número de filtros. Los bloques residuales dentro de cada bloque utilizan conexiones de salto para agregar la entrada original a la salida, facilitando el flujo de gradientes y permitiendo que la red aprenda mapeos residuales de manera efectiva. Las capas finales incluyen agrupación promedio y capas completamente conectadas con activación SoftMax para clasificación. ResNet50, una variante específica, comprende cuatro bloques con bloques residuales variables en cada uno, lo que conduce a un rendimiento de vanguardia en diversas tareas de reconocimiento visual. En la Tabla 3-1 se puede encontrar un resumen de la variante ResNet50.

Tabla 3-1: Cada bloque consta de varios bloques residuales (ResBlock) apilados juntos. El paso indica la configuración utilizada en cada bloque. El número de filtros representa el número de filtros convolucionales utilizados en cada ResBlock; la función de activación utilizada en toda la red es ReLU.

Bloque	Tamaño de salida	Capas	Paso	Número de filtros
Bloque 1	56x56x256	ResBlock1-1, ResBlock1-2, ResBlock1-3	1	64, 64, 256
Bloque 2	28x28x512	ResBlock2-1, ResBlock2-2, ResBlock2-3, ResBlock2-4	2	128, 128, 512
Bloque 3	14x14x1024	ResBlock3-1, ResBlock3-2, ResBlock3-3, ResBlock3-4, ResBlock3-5, ResBlock3-6	2	256, 256, 1024
Bloque 4	7x7x2048	ResBlock4-1, ResBlock4-2, ResBlock4-3	2	512, 512, 2048

3.2.4.2. Neck

"*Neck*" se refiere a un componente entre el *Backbone* (a menudo una red neuronal convolucional profunda) y la red *Head*. El *Neck* es responsable de procesar aún más las características extraídas por del *Backbone* y prepararlas para tareas de detección de objetos como clasificación y regresión de cuadros delimitadores.

En esta investigación, se adoptó FPN, una arquitectura ampliamente utilizada en visión por computadora, específicamente para tareas de detección de objetos y segmentación semántica [124]. FPN aborda el desafío de capturar información a múltiples escalas al crear una pirámide de mapas de características con diferentes resoluciones espaciales. Opera en los mapas de características de salida del *Backbone*.

FPN tiene dos componentes principales: rutas ascendentes y descendentes. La ruta ascendente toma los mapas de características de alta resolución del *Backbone*. Aplica capas convolucionales para reducir sus dimensiones espaciales mientras aumenta el número de canales. La ruta descendente luego toma los mapas de características de menor resolución y los sobremuestrea a través de una secuencia de operaciones de sobremuestreo y fusión. Estas características fusionadas se combinan con los mapas de características de alta resolución correspondientes de la ruta ascendente para crear una pirámide de múltiples escalas de mapas de características. Esta pirámide permite capturar detalles más precisos e información semántica de alto nivel en múltiples escalas. FPN ha demostrado ser eficaz para mejorar el rendimiento de detección y segmentación de objetos al permitir una mejor representación de características y el manejo de objetos de diversas escalas y tamaños. La Tabla 3-2 lo resume.

Tabla 3-2: Feature Pyramid Network. Estructura y componentes.

Componente	**Tamaño de salida**	**Capas**
Entrada	Mapa de características de alta resolución	-
Ruta ascendente	Mapa de características de resolución reducida	Capas convolucionales
Ruta descendente	Mapas de características sobremuestrados	Operaciones de sobremuestreo y fusión
Pirámide de características	Pirámide de características a diferentes escalas	-

3.2.4.3. Cascade R-CNN

Previamente se analizó este tipo de modelos, el cual introduce una cascada de

detectores progresivamente más precisos. En esta investigación, Cascade R-CNN logra una alta precisión de detección al reducir los falsos positivos y aumentar la tasa de recuperación mediante el empleo de este enfoque en cascada. Es un marco poderoso para tareas que requieren una localización y clasificación precisas de objetos, como en escenarios con alta densidad de objetos u objetos pequeños.

3.2.5. El problema del desbalance

Como se observó en la sección de exploración de datos, se descubrió un gran problema de desequilibrio en el conjunto de datos. En el trabajo de Oksuz et al. [129], se realizó una investigación exhaustiva relacionada con el problema de desequilibrio en la detección de objetos. Identificaron ocho problemas principales subdivididos en cuatro tipos: desbalance de clases, desbalance de escala, desbalance espacial y desbalance con la función objetivo. Los desequilibrios son:

- De clase de primer plano-fondo.
- De clase de primer plano-primer plano.
- De escala a nivel de objeto/cuadro.
- A nivel de características.
- En la pérdida de regresión.
- En la distribución de IoU.
- En la ubicación del objeto.
- De penalizaciones.

Abordar los problemas incluye manejar diferentes escenarios, que se discutirán a continuación.

3.2.5.1. Desbalance de clases

Este problema se divide en *foreground-background*, donde la mayoría de los cuadros delimitadores se etiquetan como *background* y desequilibrio de *foreground-foreground*. Centramos la solución en el primer enfoque, a través del muestreo duro, lo que significa eliminar algunos cuadros delimitadores para que no afecten el proceso de entrenamiento. El fundamento subyacente de este enfoque se basa en la noción de entrenar detectores de objetos utilizando instancias desafiantes,

que a su vez provocan valores elevados de la función de pérdida [129].

En el contexto de la detección de objetos, el muestreo duro se puede utilizar para seleccionar un subconjunto de cuadros delimitadores que representan ejemplos de baja confianza o de difícil detección. Estos ejemplos son más probables de contribuir al error de entrenamiento, por lo que enfocarse en ellos puede mejorar significativamente el rendimiento del detector de objetos.

Este trabajo aplicó OHEM [130], una técnica utilizada en el entrenamiento de redes neuronales profundas, particularmente en la detección de objetos y otras tareas similares, para enfocarse en ejemplos desafiantes o difíciles durante el proceso de entrenamiento. Aborda el problema de conjuntos de datos desequilibrados donde el número de ejemplos positivos (por ejemplo, objetos de interés) es significativamente menor que el número de ejemplos negativos (por ejemplo, regiones de *background* o no objeto).

OHEM funciona mediante la extracción selectiva de ejemplos difíciles durante cada iteración de entrenamiento. En lugar de muestrear ejemplos aleatoriamente, OHEM identifica los ejemplos más desafiantes o mal clasificados según un criterio específico, como la función de pérdida o la confianza de clasificación. Estos ejemplos difíciles, generalmente falsos negativos o ejemplos positivos difíciles, se utilizan luego para actualizar los parámetros del modelo y mejorar el rendimiento en casos desafiantes. La secuencia de pasos es la siguiente:

1. En cada iteración del conjunto de anotaciones de imagen, se obtienen puntuaciones de confianza.
2. Se seleccionan las regiones de detección con las puntuaciones más bajas.
3. Se crea un conjunto representativo más pequeño (conjunto de entrenamiento OHEM) a partir de los ejemplos difíciles.
4. El subconjunto seleccionado realiza un paso de entrenamiento adicional en el detector de objetos.

La extracción de ejemplos difíciles es una técnica eficaz para mejorar el rendimiento de los detectores de objetos, especialmente cuando se enfrentan a conjuntos de datos desequilibrados o ejemplos difíciles de detectar. Al enfocarse en los ejemplos más desafiantes, OHEM puede ayudar a que el detector de objetos aprenda a distinguir mejor entre objetos y *background*, y a mejorar su precisión general.

3.2.5.2. Desbalance en la escala

Surgen dos tipos de problemas aquí: el desequilibrio de escala a nivel de cuadro, cuando un grupo de tamaños de objeto está sobrerrepresentado, moviendo el modelo entrenado hacia esta región de interés sobrerrepresentada. Para resolver este problema, se introdujo FPN (descripción en la sección 3.2.4.3), utilizando su estructura de conexión ascendente, descendente y lateral.

El otro problema presenta un desequilibrio de escala a nivel de característica, y se realizaron algunas experimentaciones a través de *Path Aggregation Network* (PANet) [131], detalle y discusión en secciones futuras.

3.2.5.3. Desequilibrio en la distribución de IoU

El desequilibrio en la distribución de IoU se produce típicamente cuando hay un número desproporcionado de cajas delimitadoras predichas con valores de IoU bajos (lo que indica una mala localización o predicciones inexactas) en comparación con los valores de IoU altos (lo que indica predicciones precisas). Este desequilibrio puede provocar una evaluación sesgada y puede afectar el rendimiento general de los modelos de detección de objetos [129].

La intuición para resolver el problema es Cascade R-CNN, que se describe en las secciones 3.1.1.1 y 3.2.4.3. Normalmente, se utiliza un umbral de IoU igual a 0.5 en la detección de objetos, lo que produce positivos de baja calidad. Los umbrales altos pueden mejorar la calidad, pero el sobreajuste del entrenamiento será un problema debido a la desaparición del muestreo positivo. Aquí es donde Cascade R-CNN entra en acción debido a su capacidad, a través de una serie de etapas, para reducir el sobreajuste del entrenamiento y mejorar la calidad de las

propuestas de forma secuencial [129].

3.2.5.4. Desequilibrio en la pérdida de regresión

Un desequilibrio en la pérdida de regresión se refiere a un desequilibrio o desproporción en la distribución de los objetivos de regresión de los rectángulos delimitadores. En la detección de objetos, junto con la tarea de clasificación, el modelo también se entrena para regredir las coordenadas de los rectángulos delimitadores que encierran con mayor precisión los objetos de interés [129].

Un desequilibrio en la pérdida de regresión puede surgir cuando existe una diferencia significativa en el número o el nivel de dificultad de los objetos con tamaños, proporciones de aspecto o posiciones variables en el conjunto de datos. Por ejemplo, si el conjunto de datos contiene muchos objetos pequeños, pero relativamente menos objetos grandes, o si ciertas clases tienen una frecuencia mayor que otras, los objetivos de regresión asociados a estos objetos pueden estar desequilibrados, y esta es exactamente la situación con las lesiones en el conjunto de datos, una discrepancia entre las características relacionadas con el RD y las características relacionadas con el glaucoma.

Este desequilibrio en los objetivos de regresión puede conducir a un aprendizaje sesgado durante el entrenamiento, ya que el modelo puede priorizar la optimización de la pérdida de regresión para la clase mayoritaria o los objetos más fáciles de predecir, mientras que potencialmente descuida los objetos minoritarios o más desafiantes.

Las funciones de pérdida basadas en IoU que se han empleado son fundamentalmente IoU Loss [132], Bounded IoU Loss [133], GIoU Loss [134], DIoU Loss y CIoU Loss [135], pero los principales inconvenientes de estas funciones de pérdida están relacionados con la estrategia de umbral de IoU, lo que significa un mayor número de muestreadores de entrenamiento negativos, desequilibrio de escala porque a los objetos con escalas más grandes se les asignan más muestras positivas que a los objetos de escalas pequeñas y falla en la compensación de muestras.

Para resolver estos problemas expuestos, se adoptó la solución de Xu et al. [136]. Proponen una distancia de Wasserstein normalizada (NWD) con una asignación de etiquetas basada en RanKing (RKA) para gestionar la detección de objetos diminutos en el dominio de las imágenes aéreas. Aquí, se tradujo ese enfoque a la detección de lesiones retinianas, que también incluye objetos diminutos.

La intuición de los autores ha modelado el cuadro delimitador como una distribución gaussiana bidimensional, donde el píxel central cx, cy tiene el mayor peso y disminuye hacia el ancho y alto de los límites (w, h). Por lo tanto, el cuadro delimitador horizontal se representa como $R = (cx, cy, w, h)$, y su distribución gaussiana bidimensional es:

$$\mathcal{N}(\mu, \Sigma), \quad (1)$$

donde μ representa la coordenada central, y Σ representa las longitudes del semi-eje a lo largo de los ejes x y y.

$$\mu = \begin{bmatrix} cx \\ cy \end{bmatrix}, \Sigma = \begin{bmatrix} \frac{w^2}{4} & 0 \\ 0 & \frac{h^2}{4} \end{bmatrix}, \quad (2)$$

El segundo paso fue usar la distancia de Wasserstein para medir la similitud entre dos distribuciones gaussianas bidimensionales, una para cajas de ancla (ab) $\mathcal{N}_{ab}$, y la segunda para cajas de *ground truth* (gtb) $\mathcal{N}_{gtb}$.

$$W_2^2(\mathcal{N}_{ab}, \mathcal{N}_{gtb}) = \left\| \left(\left[cx_{ab}, cy_{ab}, \frac{w_{ab}}{2}, \frac{h_{ab}}{2} \right]^T, \left[cx_{gtb}, cy_{gtb}, \frac{w_{gtb}}{2}, \frac{h_{gtb}}{2} \right]^T \right) \right\|_2^2, \quad (3)$$

donde $\mathcal{N}_{ab}, \mathcal{N}_{gtb}$, la distribución gaussiana, se modela a partir del cuadro delimitador horizontal, y ||.|| es la norma de Frobenius.

Finalmente, se aplicó una función de transformación exponencial no lineal a $W_2^2(\mathcal{N}_p, \mathcal{N}_{gt})$, para utilizar esta distancia como una métrica necesaria para normalizar los valores entre (0, 1]. Entonces, la ecuación final es:

$$NWD = e\left[-\frac{\sqrt{W_2^2(\mathcal{N}_{ab},\mathcal{N}_{gtb})}}{C}\right], \quad (4)$$

C es un hiperparámetro constante. Los autores afirman que esta métrica refleja mejor la similitud entre dos distribuciones, incluso sin superponerse.

Se aplicó RKA de anclas en combinación con NWD para aumentar el rendimiento de la nueva métrica. La estrategia fue:

- Convertir los cuadros de anclaje y los cuadros de *ground truth* en una distribución gaussiana bidimensional y calcular el NWD entre sí.
- Obtener la matriz de puntuación de NWD y ordenar cada anclaje con respecto a un *ground truth* particular.
- Asignar etiquetas positivas a las anclas con la puntuación *Top k* sobre un *ground truth* particular.

Este mecanismo evita el uso de asignación basada en umbrales, el cual crea un desequilibrio para objetos de diferentes tamaños. Para un análisis más profundo, consulte el documento original [136].

3.2.6. Posprocesamiento

El posprocesamiento es un paso esencial en la detección de objetos, ya que permite el refinamiento, la interpretación y el filtrado de los resultados de detección en bruto obtenidos del algoritmo. Las razones claves por las que son importantes son el refinamiento del cuadro delimitador, la asignación de etiquetas de clase, el umbral de confianza para filtrar la detección de baja confianza, la interpretación y visualización de los resultados, y la evaluación del rendimiento.

Las técnicas de supresión no máxima son cruciales en las tareas de posprocesamiento. Ayudan a eliminar detecciones redundantes, manejan múltiples objetos, permiten un equilibrio flexible en los intercambios de Precision-Recall y proporcionan un cálculo eficiente [137].

Weighted Cluster-DIoU-NMS (WDIoUNMS) es una variante del algoritmo de supresión no máxima (NMS) comúnmente utilizado en tareas de detección de objetos en visión por computadora. El trabajo de Zheng et al. inspiró esta variante [138] e incorpora dos conceptos clave: IoU de distancia y agrupamiento ponderado.

IoU de distancia (DIoU) [135] es una métrica de evaluación que combina la superposición de IoU de dos cuadros delimitadores con una medida de distancia entre sus centros. A diferencia del IoU tradicional, DIoU considera la información de ubicación y forma de los cuadros delimitadores, lo que lo hace más efectivo para medir su similitud.

El agrupamiento ponderado es una técnica que asigna pesos a los cuadros delimitadores antes de realizar la supresión no máxima. Estos pesos otorgan más importancia a ciertos cuadros que a otros, mejorando la precisión y la calidad de las detecciones finales.

El algoritmo WDIoUNMS sigue estos pasos:

1. Ordenar los cuadros delimitadores según sus puntuaciones en orden descendente.
2. Calcular la métrica DIoU entre todos los cuadros delimitadores para obtener una matriz de distancia DIoU.
3. Aplicar una operación triu (matriz triangular superior) a la matriz de distancia DIoU para obtener una matriz IoU.
4. Iterar un número máximo de veces y actualizar la matriz IoU en función de un umbral predefinido.
5. Calcular pesos para cada cuadro delimitador utilizando una función de peso que considera la matriz IoU, las puntuaciones y una operación exponencial.
6. Pesar los cuadros delimitadores por los pesos calculados y normalizar los resultados.

7. Realizar la supresión no máxima final, conservando solo los cuadros con las puntuaciones más altas, considerando las puntuaciones ponderadas y el umbral IoU.

WDIoUNMS tiene como objetivo mejorar la calidad de las detecciones finales al considerar tanto la similitud espacial de los cuadros delimitadores como sus puntuaciones relativas, lo que puede conducir a una supresión más precisa de cuadros redundantes y la selección de las detecciones más relevantes.

4. Resultados y evaluación

En este capítulo se busca evaluar el desempeño de las metodologías antes propuestas, mostrando los resultados experimentales en la segmentación de instancias del DO y la CO; así como la detección de lesiones en la retina asociadas a la RD y el glaucoma.

4.1. Elementos de configuración

4.1.1. Parámetros e hiperparámetros

El entrenamiento de modelos de detección de objetos y arquitecturas de redes neuronales generalmente requiere la optimización de muchos parámetros interdependientes. El proceso de selección de hiperparámetros puede ser realizado por diseñadores humanos o métodos de optimización de hiperparámetros [139]. Ejemplos de métodos anteriores son Random Search [140], Grid Search [141] y Gradient-based [142]. En este caso se utilizó un ajuste manual de hiperparámetros, siguiendo enfoques bien establecidos en la literatura.

El muestreador (*Sampler*) por GPU o tamaño de lote para el resultado reportado fue de ocho, con dos trabajadores (*Workers*) por GPU y el número final de épocas fue cincuenta para la detección de lesiones, mientras que para la tarea de segmentación el número total de épocas fue de doce.

El optimizador AdamW [143] se empleó para la optimización, con una tasa de aprendizaje inicial establecida en 0.0025, y se utilizó el recocido de coseno (*Cosine*

annealing), incorporando técnicas de reinicio en caliente (*Warm restart*) [144]. Originalmente diseñado para el optimizador de descenso de gradiente estocástico (SGD), investigaciones recientes indican un mejor rendimiento cuando se aplica el recocido de coseno con AdamW [145].

Otra característica que se ha ajustado son las cajas de anclaje (*anchor boxes*), un parámetro esencial para la detección de objetos de calidad. Los ajustes de anclaje se especifican con escalas y relaciones de anclaje, mientras que los pasos de anclaje corresponden a los pasos del mapa de características. Para obtener más escalas en cada ubicación, de ahí la alta posibilidad de fijar el objeto correctamente. Se agregaron más escalas y relaciones en esta investigación. La Tabla 4-1 muestra los parámetros e hiperparámetros seleccionados.

Tabla 4-1: Parámetros e hiperparámetros ajustados durante el entrenamiento.

Hiperparámetros	**Valores**
Muestras por GPU	8
Trabajadores por GPU	2
Épocas	12/50
Optimizador	AdamW
Tasa de aprendizaje	0.0025
Horario de tasa de aprendizaje	Cosine Annealing
Redimensiones multi-scale	1333 x 640 to 1333 x 960

4.1.2. Funciones de pérdida

Las funciones de pérdida son una métrica que mide la distancia entre las predicciones de una red neuronal y los valores reales a través del cálculo de un error para cada ejemplo del conjunto de datos. Luego, estos errores se suman y se promedian para obtener un único número representativo de la distancia entre las predicciones de la red neuronal y los valores reales. Se utilizan para evaluar el rendimiento de una red neuronal durante su entrenamiento.

El objetivo del entrenamiento de una red neuronal es encontrar los parámetros de la red (pesos y sesgos) que minimicen la función de pérdida. Esto significa encontrar los parámetros que generen las predicciones más cercanas a los valores reales, convirtiéndose en un problema de optimización que busca minimizar la función de pérdida. En muchos casos, la función de pérdida no se puede resolver de forma analítica. Esto significa que no existe una fórmula matemática que permita encontrar los parámetros que la minimicen, pero si se pueden aproximar a través de estos algoritmos de optimización de forma iterativa [146].

La pérdida total en los modelos de detección de objetos es la suma de la pérdida de la clasificación, la localización y en caso de que se aplique de la segmentación. En esta investigación se realizaron tanto tareas de detección como de detección y segmentación, aplicándose así la variante correspondiente en cada caso.

Para la tarea de segmentar el disco y la copa ópticas, la ecuación de la pérdida total es la siguiente:

$$L_{Total} = L_{Cls} + L_{Reg} + L_{Mask} \qquad (5)$$

L_{Cls} es la pérdida de clasificación, la cual utiliza una función de pérdida logarítmica sobre dos clases, p_i, la probabilidad predicha, y q_i, la etiqueta del *ground truth*, ver Ecuación 6:

$$L_{CLs}(p_i, q_i) = -q_i log p_i - (1 - q_i)\log(1 - p_i) \qquad (6)$$

L_{Reg} es la pérdida de regresión de caja delimitadora. Es el error cuadrático medio que se aplica típicamente entre los puntos originales u_i y los puntos predichos v_i sobre el vector de coordenadas central, ancho y alto, $i \in [x, y, w, h]$, ver Ecuación 7:

$$L_{Reg} = MSE(u_i, v_i) \qquad (7)$$

Finalmente, L_{Mask}, emplea una función de entropía cruzada binaria promedio sobre una máscara de dimensión *m x m* asociada con la clase k del *ground truth*. Consulte la Ecuación 8 para obtener más detalles, donde $x_{i,j}$ es la etiqueta de la

celda (i, j) en el *ground truth* y $y_{i,j}^{k}$ es el valor generado por el modelo de la misma celda y clase k [42].

$$L_{Mask} = -\frac{1}{m^2}\sum_{1\leq i,j\leq m} x_{i,j} log y_{i,j}^{k} + (1 - x_{i,j})\log\left(1 - y_{i,j}^{k}\right) \quad (8)$$

Para la tarea de detección de lesiones en la retina, la función de pérdida final combina la pérdida de clasificación con la pérdida de regresión de cajas delimitadoras. Se utilizó *SmoothL1Loss* [40] para la última porque reduce el efecto de los valores atípicos y es más robusta, y es más estable para pequeños errores. En la clasificación de pérdidas, se experimentó con la conocida *Cross Entropy Loss* en combinación con OHEM y *Asymmetric Focal Loss* [147]. La función de pérdida total es la siguiente:

$$Loss = Loss_{Classification} + Loss_{BB\ Regression} \quad (9)$$

La pérdida de entropía cruzada (CE) [148] pertenece a la familia de funciones exponenciales, por lo que siempre es convexa, lo que la hace adecuada para tareas de clasificación; en esta parte de la investigación, clasificación multiclase, donde cada muestra puede pertenecer a una de las C clases. La CNN estará formada por C neuronas de salida, formando el vector s (*Scores*). El vector objetivo (*ground truth*) t será un vector *one-hot*, con una clase positiva y C-1 clases negativas. Descripción en Ecuación 10.

$$CELoss = -\sum_{i}^{C} t_i \log(s_i) \quad (10)$$

La pérdida asimétrica (ASL) se diseñó para manejar el desequilibrio de etiquetas positivas y negativas. La función de pérdida permite la ponderación dinámica descendente y el umbral duro de muestras negativas fáciles, mientras se descartan muestras potencialmente mal etiquetadas. En el artículo base se combina mecanismos de enfoque asimétrico para reducir el impacto de las muestras negativas fáciles en la función de pérdida, lo que se logra mediante un umbral suave utilizando los parámetros de enfoque γ- y γ+, y el cambio de probabilidad, que implica un umbral duro de muestras negativas directas, lo que significa que las muestras negativas se descartan por completo cuando su probabilidad es

excesivamente baja [147]. Consulte la definición en la Ecuación 11.

$$ASL = \begin{cases} L_+ = (1-p)^{\gamma_+} \log p \\ L_- = (p_m)^{\gamma_-} \log(1-p_m) \end{cases} \qquad (11)$$

donde la probabilidad desplazada se define por $p_m = \max(p - m, 0)$, p la probabilidad de salida de la red y m un factor de desplazamiento.

4.2. Métricas de evaluación

Evaluar modelos de DL es crucial por varias razones. El rendimiento de los modelos de evaluación nos permite analizar su desempeño y determinar qué tan bien resuelven el problema. Ayudan en la selección de modelos al comparar su comportamiento y seleccionar el mejor rendimiento en nuestro problema específico.

En esta investigación se trabajó con modelos de detección de objetos, por lo que sus métricas principales fueron empleadas en su evaluación. Estas son:

- Verdadero positivo (TP), la etiqueta de la muestra es positiva y se clasifica como tal. Verdadero negativo (TN), la etiqueta de la muestra es negativa y se clasifica como tal. Falso positivo (FP): la etiqueta de la muestra es negativa, pero se clasifica como positiva. Falso negativo (FN): la etiqueta de la muestra es positiva, pero se clasifica como negativa. Estos recuentos a menudo se informan como números sin procesar y se pueden usar para calcular varias otras métricas cómo las que se mencionan a continuación.
- Precisión y Recuperación (Precision/Recall): La precisión mide la fracción de objetos predichos correctamente de todos los objetos predichos, mientras que el recuerdo mide la fracción de objetos predichos correctamente de todos los objetos del *ground truth*. Estas métricas proporcionan información sobre la capacidad del modelo para detectar objetos con precisión y evitar falsos positivos (precisión) y falsos negativos (recuerdo) [149].

$$Recall = \frac{TP}{TP+FN} \qquad (12)$$

$$Precision = \frac{TP}{TP+FP} \qquad (13)$$

- F1-Score: es la media harmónica entre la recuperación y la precisión.

$$F1_{Score} = \frac{2*TP}{2*TP+FP+FN} \qquad (14)$$

- Intersección sobre Unión (IoU): IoU mide la superposición entre el cuadro delimitador predicho y el cuadro delimitador del *ground truth* de un objeto. Se calcula como la relación entre el área de intersección y el área de unión de los dos cuadros. IoU se usa ampliamente como una métrica de evaluación principal en la detección de objetos porque cuantifica la precisión de la localización.

$$IoU = \frac{TP}{TP+FP+FN} \quad (15)$$

- Precisión Promedio: AP es una métrica comúnmente utilizada para evaluar el compromiso entre precisión y recuerdo en la detección de objetos. Considera un rango de umbrales de IoU (p. ej., de 0.5 a 0.95) y calcula la precisión y el recuerdo en cada umbral. El AP es la precisión promedio en todos los umbrales, proporcionando una medida general de la calidad de detección [115].

$$AP = \sum_{i=0}^{i=n-1}(r_{i+1} - r_i)p_{interpolation}(r_{i+1}) \qquad (16)$$

Donde *r* es el número total de muestras relevantes, *n* es el número de umbrales y $p_{interpolation}$ es la precisión en cada nivel de recuerdo *r*, definida por la ecuación 17, y donde *p(r̃)* es la precisión medida en el recuerdo *r̃*, que es el recuerdo que excede *r*.

$$p_{interpolation}(r_{i+1}) = maxp(\tilde{r}), \tilde{r} \geq r_{i+1} \quad (17)$$

Según [149], estas métricas son adecuadas para la segmentación de imágenes médicas, junto con la puntuación F1, que es ligeramente diferente de IoU porque este penaliza la subsegmentación y la sobresegmentación más que F1-Score. La métrica F1-Score se usa a menudo para cuantificar el rendimiento de los métodos de segmentación de imágenes. Esta métrica es un orden de cuán similares son dos objetos.

- Precisión Promedio Media: mAP es el promedio de los valores de AP calculados para diferentes categorías de objetos en tareas de detección de objetos de múltiples clases. Proporciona una evaluación integral del rendimiento del modelo en múltiples clases. A menudo se informa el mAP en un umbral de IoU específico (por ejemplo, 0.5 o 0.75).

Dado que la presente investigación se basa en el formato COCO, se adoptó su métrica de evaluación, que incluye el análisis previo [114]. Además, hemos utilizado un kit de herramientas completo conocido como *Toolbox for Identifying Object Detection Errors* (TIDE), que clasifica los errores en seis tipos distintos: Error de clasificación, Error de localización, Errores en la clasificación y localización, Error de detección duplicada, Error de fondo y Error de *ground truth* omitido [150]. Además, introducen un enfoque para medir la contribución de cada error, aislando su efecto en el rendimiento general.

4.3. Experimentación y resultados en la segmentación de instancias.

La experimentación comienza con una fracción del conjunto de datos REFUGE, específicamente 100 imágenes para entrenamiento, 30 para validación y 30 para la prueba. Este subconjunto se tomó, debido a la engorrosa naturaleza del proceso de etiquetado, para observar una primera aproximación del comportamiento de los modelos a evaluar. También nos da una idea de cómo se comporta el modelo con un número limitado de imágenes. La tarea se desarrolló sin escala múltiple (WM) y escala múltiple (MS).

Se reportan tres criterios: AP[IoU=0.50:0.95], donde el AP se promedia sobre múltiples valores de IoU, lo que recompensa a los detectores con mejor localización. Esta se tomará como la métrica principal; también se reportan (AP) [IoU=0.50] y (AP) [IoU=0.75] ya que son más comunes en la literatura y los resultados son más ajustados. Tres modelos mejoran su rendimiento con escala múltiple: GCNet, MS-RCNN y Point_Rend. El mejor resultado general fue Mask-RCNN con AP[IoU=0.50:0.95] con 0.671. Los resultados se pueden ver en la Tabla

4-2.

Tabla 4-2: Resultados de precisión media en el conjunto de datos reducido REFUGE.

Modelo	(AP)[IoU=0.50:0.95]		(AP)[IoU=0.5]		(AP)[IoU=0.75]	
	WM	MS	WM	MS	WM	MS
CARAFE	0.657	0.607	0.979	0.965	**0.771**	0.621
Cascade Mask-RCNN	0.618	0.608	**1.000**	**0.980**	0.661	0.646
SOLO	0.555	0.530	0.886	0.886	0.613	0.586
GCNET	0.584	0.595	0.980	0.960	0.608	0.638
MASK-RCNN	**0.671**	0.616	**1.000**	0.962	0.743	0.635
MS-RCNN	0.604	**0.627**	0.980	0.978	0.649	**0.676**
POINT_REND	0.582	0.607	**1.000**	0.965	0.564	0.621

Posteriormente, el experimento se repitió bajo las mismas condiciones con el conjunto de datos REFUGE completo, utilizando 400 imágenes para entrenamiento, 200 para validación y 200 para pruebas. Los resultados se pueden observar en la Tabla 4-3.

Tabla 4-3: Resultados de precisión media en el conjunto de datos completo de REFUGE.

Modelos	(AP)[IoU=0.50:0.95]		(AP)[IoU=0.50]		(AP)[IoU=0.75]		F1-Score
	WM	MS	WM	MS	WM	MS	
CARAFE	0.650	0.636	0.990	0.995	0.710	0.685	**1.0**
Cascade Mask-RCNN	0.644	**0.661**	0.985	0.990	0.716	**0.739**	0.997
SOLO	0.610	0.647	0.989	0.984	0.676	0.703	**1.0**
GCNET	0.631	0.656	0.990	0.995	0.712	0.729	**1.0**
MASK-RCNN	0.595	0.629	0.948	0.988	0.662	0.701	**1.0**
MS-RCNN	**0.654**	0.658	**0.995**	**1.000**	**0.766**	0.738	**1.0**
POINT_REND	0.632	**0.661**	0.990	0.994	0.670	0.735	**1.0**

A excepción de CARAFE, todos los modelos mejoraron su rendimiento con el enfoque de escala múltiple. Los experimentos se realizaron en el conjunto de datos G1020, con escala múltiple, ya que este enfoque muestra mejores resultados. Los resultados se muestran en la Tabla 4-4.

Se proporcionan curvas de *precision-recall* (PR) para una mejor comprensión. El primer gráfico, consulte la Figura 4-1, muestra que todos los modelos funcionan de manera excelente, lo que significa que la recuperación aumenta en cierta cantidad y la precisión no cambia; por lo tanto, todos los recuperados son verdaderos

positivos.

Tabla 4-4: Resultados de precisión media en el conjunto de datos G1020.

Model Architecture	(AP)[IoU=0.50:0.95]	(AP)[IoU=0.50]	(AP)[IoU=0.75]	F1-score
	MS	MS	MS	
CARAFE	0.624	0.948	0.632	0.963
Cascade Mask-RCNN	0.631	0.947	0.662	0.963
SOLO	0.568	0.909	0.583	0.916
GCNET	0.628	0.943	0.646	0.957
MASK-RCNN	0.613	0.941	0.621	0.963
MS-RCNN	**0.638**	0.944	**0.664**	0.963
POINT_REND	0.617	**0.956**	0.648	**0.969**

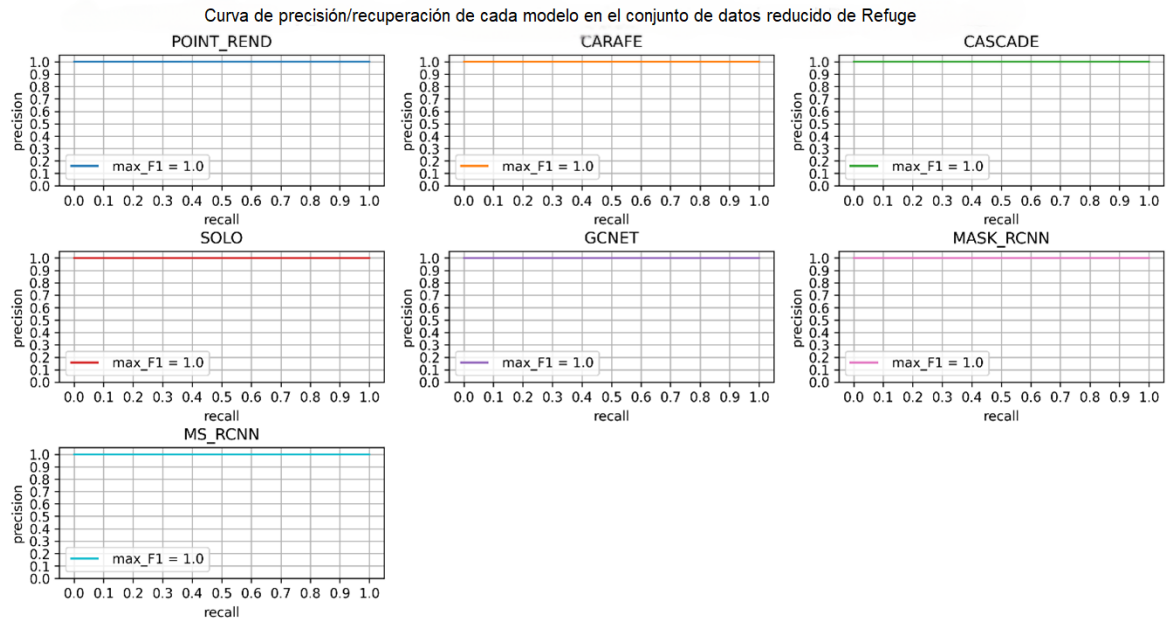


Figura 4-1: Curva de PR de cada modelo en el conjunto de datos reducido REFUGE. Se proporciona la puntuación F1-Score de cada modelo.

En la Figura 4-2, al igual que en la anterior, todos los rendimientos de los modelos son perfectos excepto Cascade Mask-RCNN. Este modelo no puede recuperar todos los verdaderos positivos.

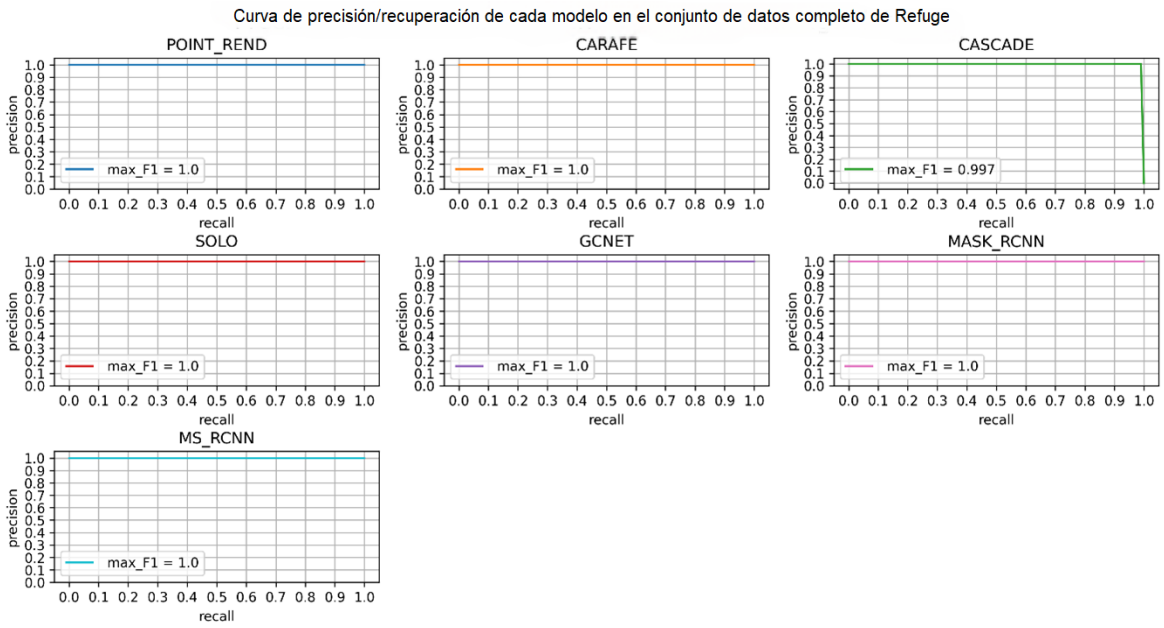


Figura 4-2: Curva de PR de cada modelo en el conjunto de datos REFUGE. Se proporciona la puntuación F1-Score de cada modelo.

La Figura 4-3 también muestra resultados prometedores. Sin embargo, este gráfico muestra implícitamente la presencia de falsos positivos asociados con errores de localización, confusión de clases y falsos negativos. Estos errores se deben a que el conjunto de datos G1020 presenta una alta diversidad en sus imágenes en comparación con el conjunto de datos REFUGE.

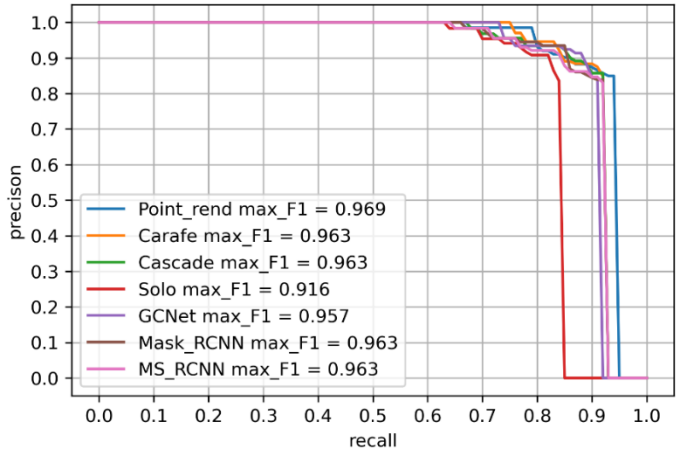


Figura 4-3: Curvas de PR de cada modelo en el conjunto de datos G1020. Se proporciona la puntuación F1-Score de cada modelo.

Se proporciona una serie de curvas de PR para cada clase con fines de

interpretación en la Figura 4-4. Se garantiza que cada curva de PR sea más alta que la anterior a medida que el entorno de evaluación se vuelve más permisivo con respecto al umbral de IoU. La leyenda se describe a continuación, con el significado de cada curva [151].

1. C75: el área bajo la curva corresponde a la métrica AP[IoU=0.75].
2. C50: el área bajo la curva corresponde a la métrica AP[IoU=0.50].
3. Loc: se ignoran los errores de localización, pero no las detecciones duplicadas.
4. Sim: PR después de eliminar los falsos positivos (fp) de supercategoría.
5. Oth: PR después de eliminar todas las confusiones de clase.
6. BG: PR después de eliminar todos los fps de fondo (y confusión de clase).
7. FN: PR después de eliminar todos los errores restantes (trivialmente AP=1).

Se presenta una explicación de la Figura 4-2, donde la curva de PR no fue perfecta para el modelo Cascade Mask-RCNN en el conjunto de datos REFUGE, mediante un análisis por clase. La Figura 4-4 muestra una línea delgada al final del gráfico que exhibe la presencia del falso negativo.

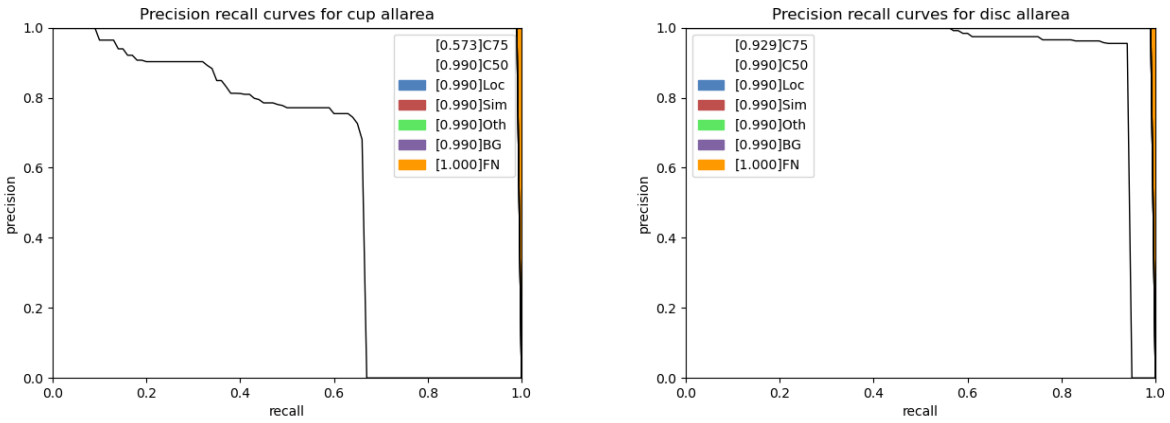


Figura 4-4: Curvas de PR por clases en Cascade Mask-RCNN sobre el conjunto de datos REFUGE. La imagen de la izquierda representa el área de la copa y la de la derecha el área del disco. Ambas muestran la presencia de falsos positivos.

Una inspección visual siempre es recomendable. En las Figuras 4-5 y 4-6 se pueden ver los resultados de la segmentación con el modelo MS-RCNN.

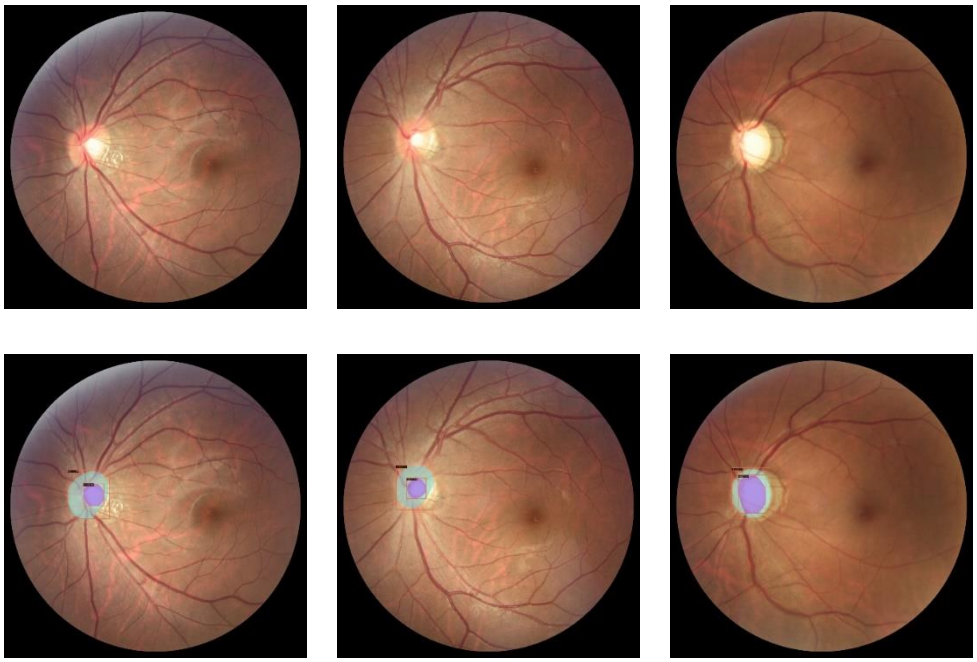


Figura 4-5: Resultado de MS-RCNN en algunas imágenes del conjunto de datos de prueba REFUGE. Primera fila de imágenes originales, segunda fila de imágenes segmentadas.

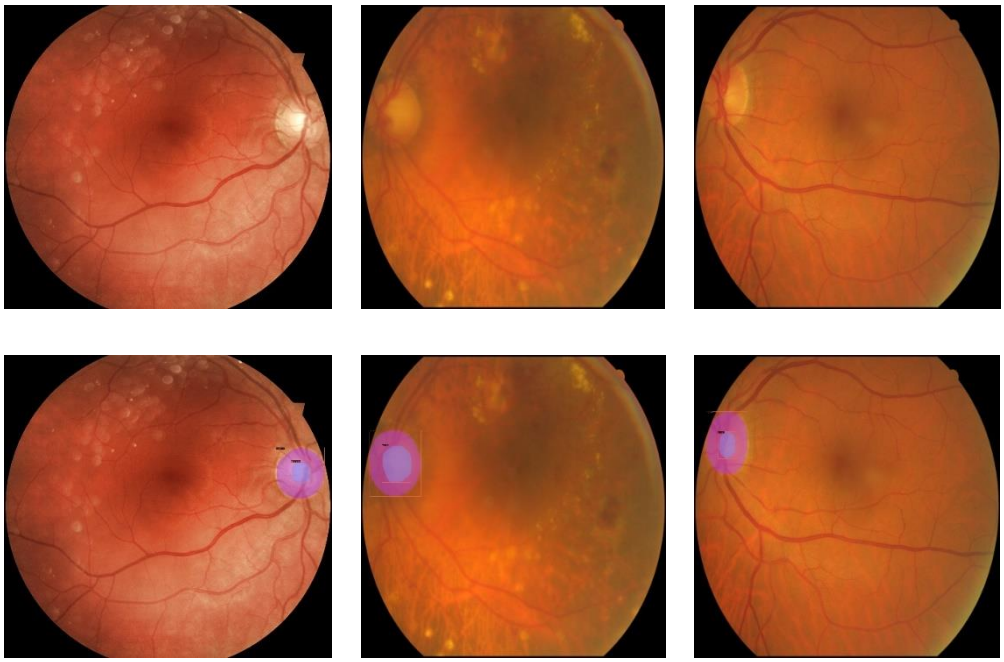
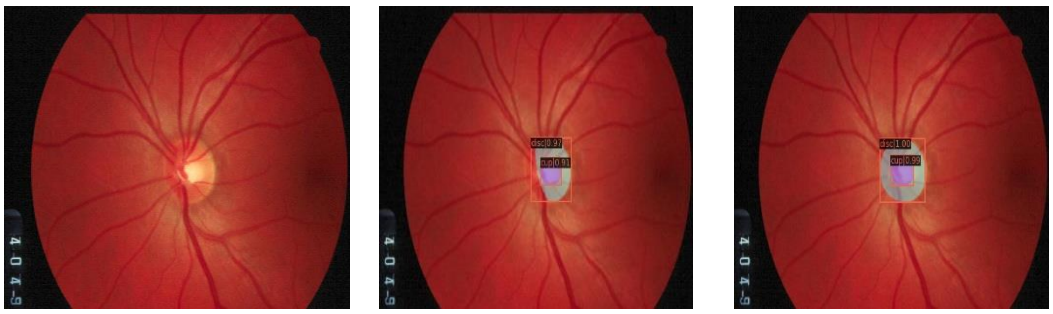


Figura 4-6: Resultado de MS-RCNN en algunas imágenes del conjunto de datos de prueba G1020. Primera fila de imágenes originales, segunda fila de imágenes segmentadas.

Los modelos entrenados con el conjunto de datos G1020 de múltiples experimentos arrojaron mejores resultados que el conjunto de datos REFUGE cuando los modelos entrenados se aplicaron a nuevos conjuntos de datos, debido a una mayor

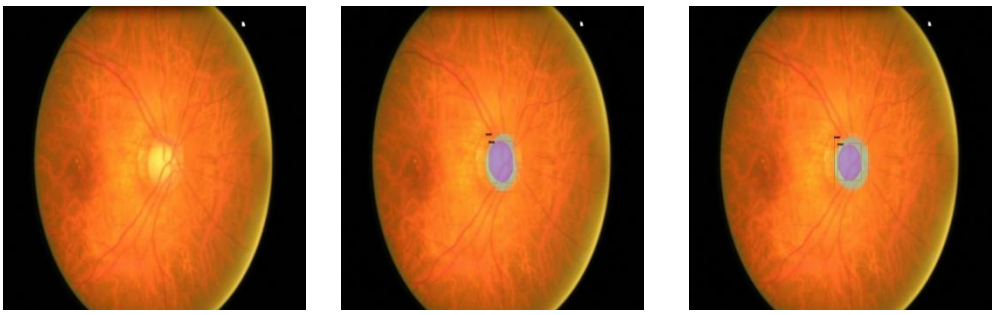
diversidad en sus imágenes. La prueba se realizó en múltiples imágenes; con fines ilustrativos, se mostraron una imagen de DRIONS-DB [152] y ORIGA DB [52] en las Figuras 4-7 y 4-8 respectivamente, con el modelo Cascade Mask R-CNN.



(a) (b) (c)

Figura 4-7: Segmentación en un conjunto de datos externo con el modelo Cascade Mask-RCNN. Imagen original de DRIONS-DB a). Resultado de la segmentación del modelo entrenado con el conjunto de datos Refuge b). Resultado de la segmentación del modelo entrenado con el conjunto de datos G1020 c).

Como se observa en la figura anterior, la segmentación obtenida para el DO no cubre toda el área esperada con los modelos entrenados en el conjunto de datos REFUGE, vea la Figura 4-7 b). Este resultado se puede ver cuando la predicción se realiza con el modelo entrenado en el conjunto de datos G1020, vea la Figura 4-7 c). En la siguiente Figura 4-8, también se puede observar la degradación de la segmentación del DO, comparando los modelos entrenados en los conjuntos de datos REFUGE y G1020.



a) b) c)

Figura 4-8: Segmentación en un conjunto de datos externo con el modelo Cascade Mask-RCNN. Imagen original de ORIGA-DB a). Resultado de la segmentación del modelo entrenado con el conjunto de datos Refuge b). Resultado de la segmentación del modelo entrenado con el conjunto de datos G1020.

4.4. Experimentación y resultados en la detección de lesiones.

Antes de presentar los resultados, se desarrollaron múltiples experimentos, desde el cambio de tamaño de imagen fijo (1333 x 800) hasta el cambio de tamaño de imagen a través de escala múltiple, un esquema de aumento simple basado en el entrenamiento con tamaños entre 1333 x 640 y 1333 x 960. Para resolver el desequilibrio de clases, dos enfoques son el muestreo duro (OHEM) y el muestreo suave con ASL. La experimentación se realizó a través de FPN y *Path Aggregation Network* (PAFPN) [131] para el desequilibrio de escala, y el posprocesamiento se realizó con y sin WDIoUNMS. Todos los experimentos se realizaron con NWD y RKA.

La distribución de imágenes para entrenamiento, validación y pruebas fue de 383, 149 y 225 respectivamente, con un total de 757 imágenes. Inicialmente, el tamaño del lote se estableció en dos y el número de épocas se estableció en doce. La evaluación de la investigación empleó principalmente mAP, ya que la tarea implicaba la detección de objetos. La mAP es una métrica predominante para evaluar la precisión de algoritmos de DL en el contexto de la detección de objetos. Siguiendo las métricas de evaluación de COCO, se informaron AP@[IoU = 0.50], ampliamente utilizado en la literatura, así como AP@[IoU = 0.50:0.95], donde AP se promedia sobre varios valores de IoU, lo que recompensa a los detectores con una mejor localización.

Además, se utilizó la curva de PR para un análisis en detalle, con una interpretación mejor adaptada sobre *Receiver Operating Characteristics* (ROC) para problemas de clases desequilibradas, y el kit de herramientas TIDE para aislar la contribución del error en lugar de solo mirar mAP. Los primeros resultados se pueden ver en la Tabla 4-5. Todas las métricas informadas se basan en conjuntos de prueba, ya que los datos no se ven en el proceso de entrenamiento como los conjuntos de entrenamiento y validación, lo que evita un posible sobreajuste en las métricas informadas.

Tabla 4-5: Resultados de la experimentación para doce épocas y dos de tamaño de lote.

Experimentos	**AP@ [IoU = 0.50]**
CE+(1333-800)	0.4295
CE+OHEM+(1333-800)	0.4331
CE+OHEM+MS	0.4358
CE+OHEM+MS+PAFPN	0.432
CE+OHEM+MS+WDIoUNMS	**0.446**
ASL+MS	0.432
ASL+MS+ PAFPN	0.435
ASL+MS+PAFPN+WDIoUNMS	0.432
ASL+MS+WDIoUNMS	**0.436**

Se pueden identificar dos enfoques principales a partir de los resultados de la Tabla 4-5, uno basado en CE con muestreo duro (OHEM) y otro con ASL como mecanismo de muestreo suave. El uso de multiescala demostró una ligera mejora, por lo que se adoptó para el resto de los experimentos. Cuando no se especificó PAFPN, se utilizó FPN, y este mecanismo en el cuello del modelo de detección de objetos proporciona los mejores resultados en combinación con WDIoUNMS como enfoque de posprocesamiento. Los mejores resultados se destacaron en negrita.

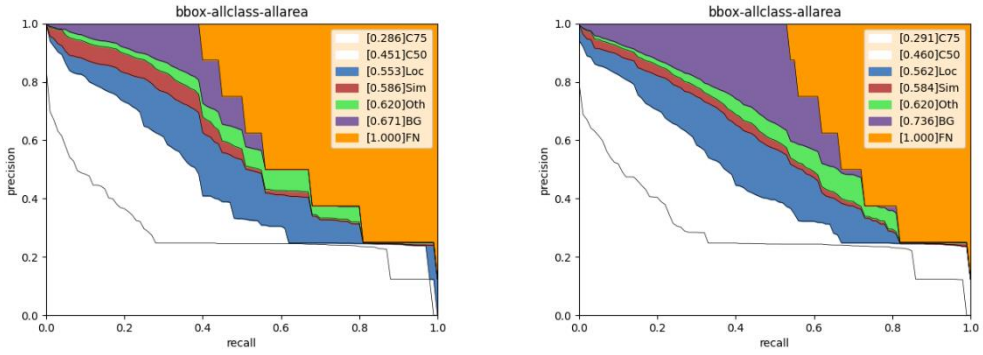
Tras el desarrollo de estos nuevos experimentos, se produjo un aumento en el número de épocas de doce a cincuenta, junto con una escalada en el tamaño del lote de dos a ocho. Este ajuste ha facilitado que el algoritmo ejecute actualizaciones de gradiente consistentes, al tiempo que fomenta la anticipación de mejoras en los resultados predictivos. En esta ocasión, los mejores resultados anteriores se compararon con el estado del arte, véase la Tabla 4-6.

En base a los resultados presentados en la Tabla 4-6, el presente estudio demuestra un rendimiento superior en comparación con investigaciones previas, logrando una mejora mínima del doble. Esto subraya la eficacia de aumentar el modelo de detección de objetos Cascade RCNN con la integración de las técnicas de NWD y RKA como marco fundamental. El experimento con PAFPN muestra efectividad para objetos pequeños pero peor comportamiento para todas las clases. El mejor resultado se obtuvo a través de la pérdida asimétrica con un mAP de 0.460. Es importante tener en cuenta que, en las métricas de evaluación de COCO, mAP y AP se utilizan indistintamente como la misma métrica; por lo tanto, ambas tienen el mismo significado.

Tabla 4-6: Resultados de la experimentación para cincuenta épocas y ocho de tamaño de lote. Comparación con el estado del arte.

Experimentos	**AP@[IoU= 0.50]**	**AP@[IoU=0.50:0.95] area=all**
Tao Li et. al. [54]		-
SSD	0.0015	
YOLO	0.0030	
Faster RCNN	0.0009	
Santos et. al. [103]	0.1540	-
Wejdan et. al. [106]	0.216	-
CE+OHEM+MS+WDIoUNMS	0.451	0.287
ASL+MS+WDIoUNMS	**0.460**	**0.293**

Para facilitar una mejor interpretación, se ha presentado una serie de curvas de PR. En estas curvas se superpone el impacto acumulativo de los errores individuales, estableciendo una relación. La Figura 4-9 muestra las curvas de PR para CE+OHEM+MS+WDIoUNMS y ASL+MS+WDIoUNMS.



a) CE+OHEM+MS+WDIoUNMS b) ASL+MS+WDIoUNMS

Figura 4-9: Curvas de PR para CE+OHEM+MS+WDIoUNMS y ASL+MS+WDIoUNMS.

En la curva de PR, el eje x corresponde a la *Recall*, mientras que el eje y corresponde a la *Precision*. Esta curva es significativa cuando se manejan clases desequilibradas, ya que evalúa principalmente el rendimiento de las clases positivas. En consecuencia, el objetivo en el espacio PR es ocupar la esquina

superior derecha (1, 1), lo que significa que el predictor identificó correctamente todas las instancias positivas (*Recall* = 1) y que cada clasificación positiva fue efectivamente precisa (*Precision* = 1).

A partir de las figuras anteriores, los resultados están lejos del objetivo principal, la esquina superior derecha. Una primera interpretación se puede tomar de las áreas de error, identificando muchos falsos negativos. Sin embargo, en un análisis profundo, Bolya et al. [150], basándose en gráficos anteriores, fusionan Sim y Oth en un solo error de clasificación y crean un nuevo tipo de error que combina el error de localización y clasificación, como ambos mal localizados y clasificados. Luego, al intercambiar el orden de las etiquetas de *background* y clasificación, calculan el error. Al calcular inicialmente el error de *background*, hay una reducción notable en su impacto general, lo que sugiere que el significado real de los errores de *background* es mucho menor que el informado en las evaluaciones COCO.

La Tabla 4-7 muestra la contribución de cada error, proporcionando un análisis intuitivo de la eficacia de la configuración de los modelos, los parámetros y los hiperparámetros elegidos en esta investigación. Es imposible extraer esta información y la interpretación posterior del mAP previamente informado.

Tabla 4-7: Contribución de cada error. Exp_1= CE+OHEM+MS+WDIoUNMS, Exp_2= ASL+MS+WDIoUNMS, E=Error (clasificación, localización, ambos clasificación+localización, duplicado, background, GT perdido).

# Exp	$E_{Cls}\downarrow$	$E_{Loc}\downarrow$	$E_{Both}\downarrow$	$E_{Dupe}\downarrow$	$E_{Bkg}\downarrow$	$E_{Miss}\downarrow$
Exp_1	2.73	9.25	**1.26**	0.00	**2.04**	**15.20**
Exp_2	**1.81**	**7.92**	1.28	0.00	2.34	15.95

El impacto de la función de pérdida asimétrica para minimizar los errores relacionados con la clasificación y la localización se debe a que el mecanismo de enfoque asimétrico enfatiza la contribución de las muestras positivas y reduce la contribución de las muestras negativas a la pérdida. Además, el uso de un mecanismo de desplazamiento de probabilidad que aplica un umbral rígido ignora las muestras negativas si su probabilidad cae por debajo de un umbral bajo determinado. Este enfoque ayuda a mitigar los errores de falsos positivos.

Por el contrario, la pérdida de CE con OHEM se desempeña mejor en la

reducción del impacto de los errores relacionados con el *background* (*background* detectado como primer plano), la omisión del *ground truth* (todo el *ground truth* no detectada que no se maneja mediante errores de clasificación y localización) y los errores del tipo Ambos (clasificación y localización simultáneamente). Una justificación supone que OHEM considera todas las ROI dentro de una imagen y elige ejemplos desafiantes para el entrenamiento. Mientras tanto, los ejemplos fáciles tendrían una pérdida baja y una contribución mínima al gradiente. En consecuencia, el proceso de entrenamiento priorizaría y enfatizaría naturalmente los ejemplos difíciles, reduciendo así los errores Ambos, la omisión del *ground truth* (objetos no detectados) y las instancias en las que el *background* se detecta erróneamente como primer plano. La Figura 4-10 provee interpretación visual de la distribución de los errores.

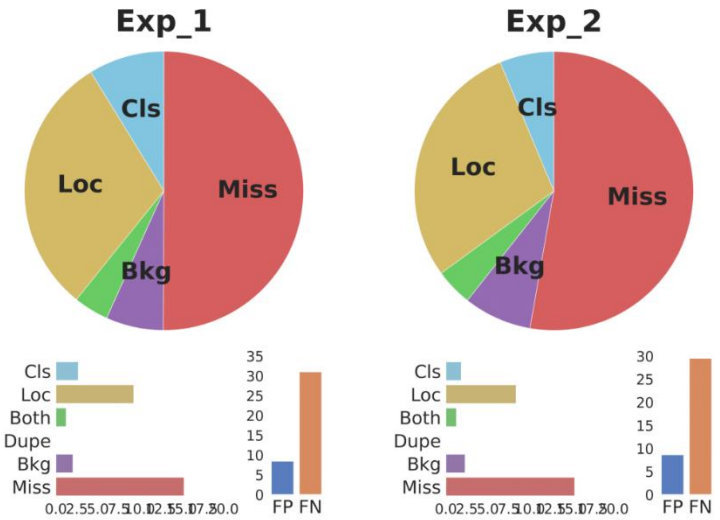


Figura 4-10: Interpretación visual de los errores en el conjunto de datos DDR. Análisis de errores específicos del modelo aplicados a varios detectores de objetos, representados mediante un gráfico circular que ilustra la contribución relativa de cada error y gráficos de barras que muestran su contribución absoluta.

Una matriz de confusión brinda más información sobre los errores y qué clase fue más difícil de detectar. Una matriz de confusión es una matriz numérica que revela las áreas de confusión de un modelo. Presenta un desglose del desempeño predictivo de un modelo de clasificación por clase, lo que nos permite comprender dónde comete errores el modelo. La matriz de confusión proporciona una representación organizada de cómo las predicciones del modelo se alinean con las

clases de datos reales.

Podemos comparar sus fortalezas y debilidades calculando una matriz de confusión para el mismo conjunto de prueba utilizando diferentes clasificadores. Tal análisis ayuda a determinar el enfoque óptimo para aprovechar múltiples clasificadores de manera efectiva. La matriz de confusión obtenida al entrenar un clasificador y evaluar el modelo entrenado en este conjunto de prueba se muestra en las Figuras 4-11 y 4-12.

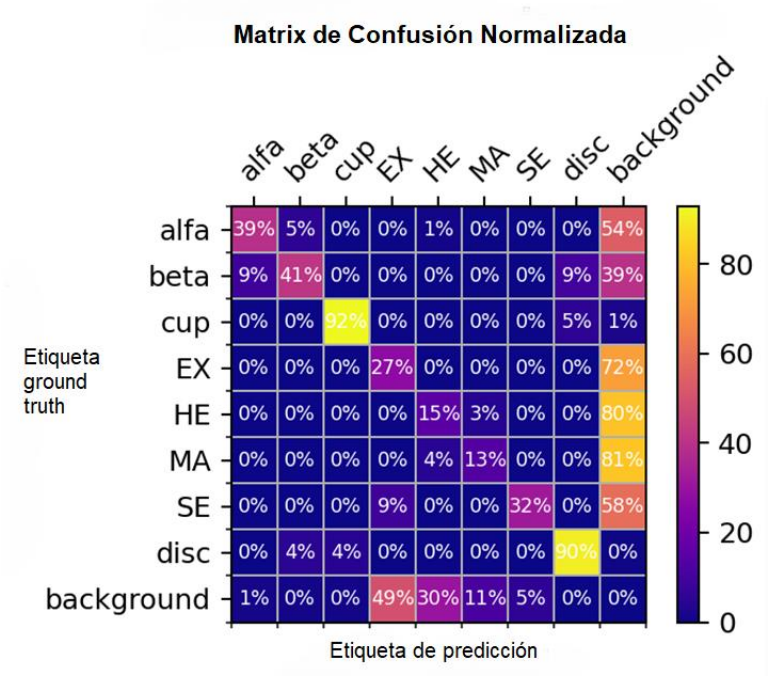


Figura 4-11: Matrix de confusión para el experimento CE+OHEM+MS+WDIoUNMS.

La diagonal principal se interpreta como la predicción correcta; mientras tanto, los errores de predicción se pueden ver fuera de ella. Los mejores resultados fueron para la CO y el DO, con un 92% y un 90%, respectivamente. Peores fueron las HE y los MA, con un 15% y un 13% de precisiones, respectivamente. Estos resultados se correlacionan con la cantidad de falso positivo de *background* (última columna), que son objetos que no pertenecen a ninguna de las clases, pero se detectan como una de ellas, y los valores fueron del 80% y 81%, respectivamente.

La siguiente matriz de confusión corresponde al experimento con ASL en la Figura 4-12. En este caso, si bien la CO y el DO, las HE y los MA siguen siendo los

mejores y los peores, respectivamente, el número de aciertos fue mayor, mejorando todas las clases relacionadas con áreas pequeñas, específicamente con MA con un valor del 21%. El falso positivo de *background* mejoró ligeramente, pero, en consecuencia, el falso negativo de *background* (en la última fila), objetos que el detector no detectó, obtuvo peores resultados, fundamentalmente con EX, con un 65% de clasificación errónea.

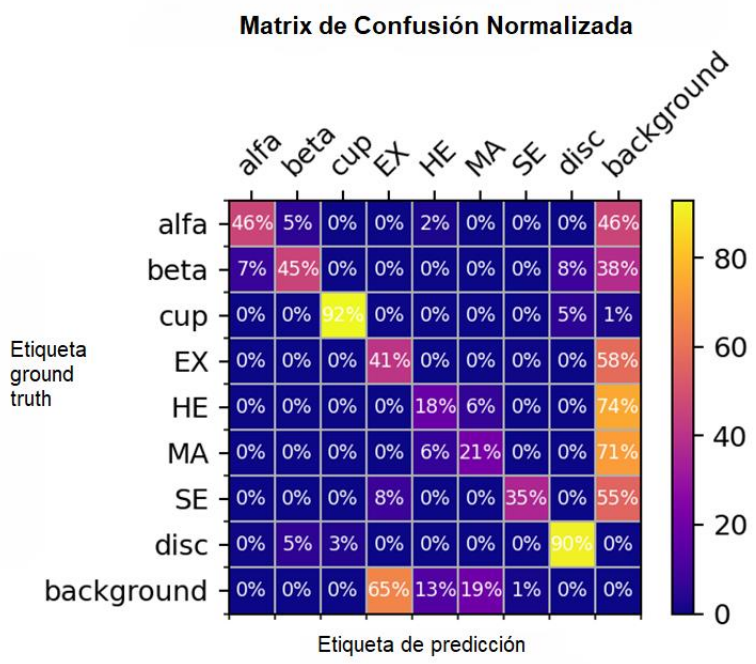


Figura 4-12: Matrix de confusión para el experimento ASL+MS+WDIoUNMS.

A partir de la matriz de confusión, es evidente que las áreas más pequeñas conducen a la peor clasificación. Los objetos pequeños tienen menos píxeles, lo que significa que hay menos información para que el modelo de detección de objetos trabaje. Los objetos diminutos se oscurecen más fácilmente por objetos más prominentes, quedando detrás de los más grandes, y estos objetos pequeños tienen más probabilidades de estar borrosos o tener un contraste bajo, lo que dificulta que el modelo distinga el objeto de su entorno.

Para inspección visual, consulte las Figuras 4-13 y 4-14 del conjunto de prueba del dataset DDR, donde además de las clases originales, ahora se pueden identificar clases como el DO, la CO y las APP. Esta imagen pertenece a

experimentos con ASL.

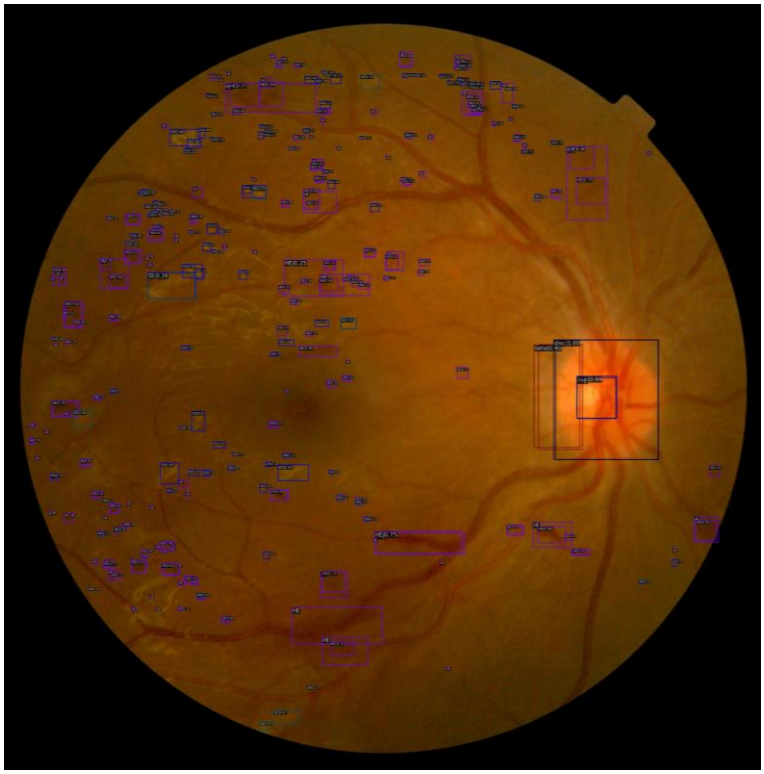


Figura 4-13: Predicción de lesiones en imágenes de retina con alta densidad de daños. Conjunto de pruebas DDR.

La imagen anterior es "007-3567-200.jpg" en el conjunto de prueba del dataset DDR. Como se puede ver, se detectaron muchas lesiones. La imagen tiene una cantidad significativa de HE, la mayoría de las cuales se detectaron. Se detectó un cuadro probabilístico para atrofia beta, copa y disco. Recuerde que estas lesiones no pertenecen al dataset DDR; su ubicación es un paso importante. Sin embargo, algunos artefactos del fondo se clasificaron como lesiones cuando no lo eran. Estos falsos positivos se clasificaron erróneamente como EX y SE principalmente. Una causa podría ser el color, que es como la lesión original; algunos MA menores también se clasificaron erróneamente.

Se puede extraer una interpretación similar de la imagen "20170505181200498.jpg" en la Figura 4-14. Se identificaron correctamente la presencia de atrofia beta, copa, disco y la multitud de EX. Se clasificaron erróneamente tres HE y dos MA.

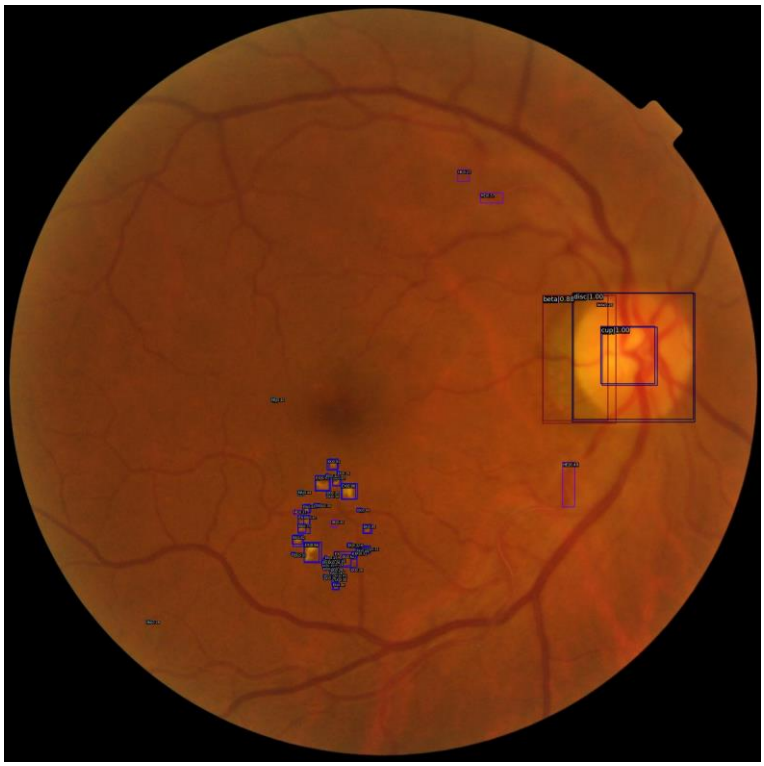


Figura 4-14: Predicción de lesiones en imagen retiniana. Conjunto de pruebas DDR.

Estos resultados sugieren que, en combinación con especialistas y médicos, las predicciones obtenidas pueden ser útiles y aliviar la carga de lectura e interpretación a la que están sometidos los médicos. También muestra la necesidad de seguir refinando el trabajo de las predicciones para disminuir el número de falsos positivos y falsos negativos.

4.5. Plataforma de software

Se han logrado resultados sorprendentemente prometedores con los sistemas de DL (DL) en los últimos años. Muchos de estos logros se han alcanzado en entornos académicos o por grandes empresas de tecnología con grupos de investigación altamente cualificados e infraestructuras de apoyo avanzadas. Para las empresas o centros académicos que no tienen grandes grupos de investigación o infraestructuras avanzadas, ha resultado difícil construir sistemas de producción de alta calidad con componentes de DL. Existe una clara falta de herramientas y prácticas recomendadas que funcionen bien para crear sistemas de DL [153].

El ML, especialmente el DL, difiere en parte de la ingeniería de software tradicional (SE) en que su comportamiento depende en gran medida de los datos del mundo exterior. De hecho, es en estas situaciones en las que el ML se vuelve útil. Una diferencia clave entre los sistemas de ML y los sistemas que no son de ML es que los datos reemplazan en parte al código en un sistema de ML, y se utiliza un algoritmo de aprendizaje para identificar automáticamente patrones en los datos en lugar de escribir reglas codificadas en forma rígida.

Para poner en producción un modelo de ML/DL, normalmente se requiere la colaboración de muchos equipos diferentes con diferentes ideas, prioridades y valores culturales. Esto no solo introduce desafíos organizativos desde un punto de vista cultural, sino también en poder estimar adecuadamente la cantidad de esfuerzo que necesitan los diferentes tipos de equipos [153].

En el trabajo de [154] se hace una descripción de las distinciones entre los procesos de Desarrollo de Software de DL y el Despliegue de Software de DL.

- Desarrollo de Software de DL: los desarrolladores utilizan potentes marcos de trabajo como *Pytorch* o *Tensorflow*. En un modelo de DL se utilizan funciones de transformación de múltiples capas para convertir entradas en salidas, y cada capa aprende un nivel superior de abstracción en los datos. Por último, los datos de prueba, que son diferentes de los datos de entrenamiento, se utilizan para ajustar el modelo.
- Despliegue de Software de DL: El proceso de despliegue se centra en la adaptación de la plataforma, es decir, en adaptar el software DL a la plataforma de despliegue. La forma más popular de despliegue es en el servidor o en plataformas en la nube. Esto permite a los desarrolladores invocar servicios basados en técnicas de DL mediante la simple llamada a un punto de acceso de la *Application Programming Interfaces* API. Los principales desafíos del despliegue de software de DL están relacionados con la adaptación a la plataforma, la optimización del rendimiento y la seguridad.

Luego de pruebas rigurosas y optimización de los modelos de DL, estos se preparan para su despliegue. Sin embargo, estos requieren de un entorno de implementación que contenga todos los recursos de hardware y los datos necesarios para que el modelo funcione de manera óptima.

4.5.1. Microservicios

Para darle solución al problema se ha trabajado con una arquitectura basada en Microservicios, la cual nace como una alternativa a la tradicional arquitectura monolítica en el desarrollo de software. La arquitectura de microservicios funciona con un conjunto de pequeños servicios que se ejecutan de manera autónoma e independiente, cada uno responsable de una funcionalidad específica y que se comunican entre sí a través de API. Entre las ventajas que tienen se encuentran la escalabilidad, una implementación sencilla, código reutilizable, agilidad en cambios, aplicación independiente y menor riesgo [155].

Los microservicios se encuentran desplegados en un servidor físico, al cual se accede desde internet, a través de los diferentes dispositivos de los usuarios y dicha comunicación pasa por un servidor proxy inverso, el cuál proporciona seguridad, equilibrio de carga y facilidad de mantenimiento. Ver figura 4-15 a continuación:

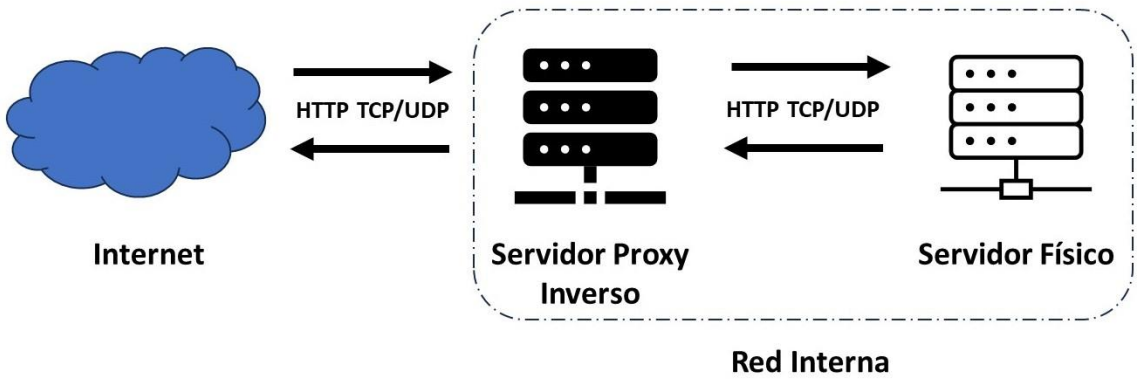


Figura 4-15: Comunicación externa/interna a través de un servidor proxy inverso, el cuál añade una capa extra de seguridad.

Los microservicios propuestos son:

- Servidor ReactJs (UI): Este servidor se encarga de la interfaz de usuario

y la presentación de los resultados del modelo de detección de objetos.

- Servidor de Token Auth (NodeJS): Este servidor se encarga de la autenticación de los usuarios y la gestión de los tokens de acceso.
- Servidor Web (NodeJS): Este servidor se encarga de la gestión de las peticiones HTTP y la comunicación con los microservicios.
- API: Este microservicio se encarga de la comunicación entre los diferentes microservicios y el modelo de detección de objetos.
- Web Sockets: Este microservicio se encarga de la comunicación en tiempo real entre el servidor y los dispositivos electrónicos.
- Servidor Deep Learning (Python): Este microservicio se encarga de la ejecución del modelo de detección de objetos.

4.5.2. Contenedor Docker

En este contexto la herramienta Docker se utiliza para trabajar con los microservicios. Docker es una plataforma de contenedores que permite a los desarrolladores empaquetar aplicaciones y servicios en un contenedor portátil y ligero. Los contenedores son reproducibles, predecibles y fáciles de modificar y actualizar, lo que facilita la colaboración entre ingenieros. Los contenedores abarcan todo el hardware, las configuraciones y las dependencias necesarias para implementar el modelo, lo que mejora la coherencia entre los equipos de ML [156].

Existen cuatro formas de desplegar modelos en producción que son modo de predicción bajo demanda, predicción por lotes, implementación en dispositivos perimetrales como modelos integrados, y la implementación mediante un servicio web, que fue precisamente la opción tomada en esta investigación.

Este método es el más simple e implementa el modelo como un servicio web, mediante la creación de una API REST y el uso de la API en aplicaciones móviles o web para los usuarios. La implementación como servicios web sirve principalmente a equipos de ML con múltiples interfaces como web, móvil y de escritorio. Las tecnologías estándar que impulsan los modelos de predicción de servicios web incluyen funciones de nube de *AWS Lambda* y Google, contenedores

Docker o portátiles como *Databricks*.

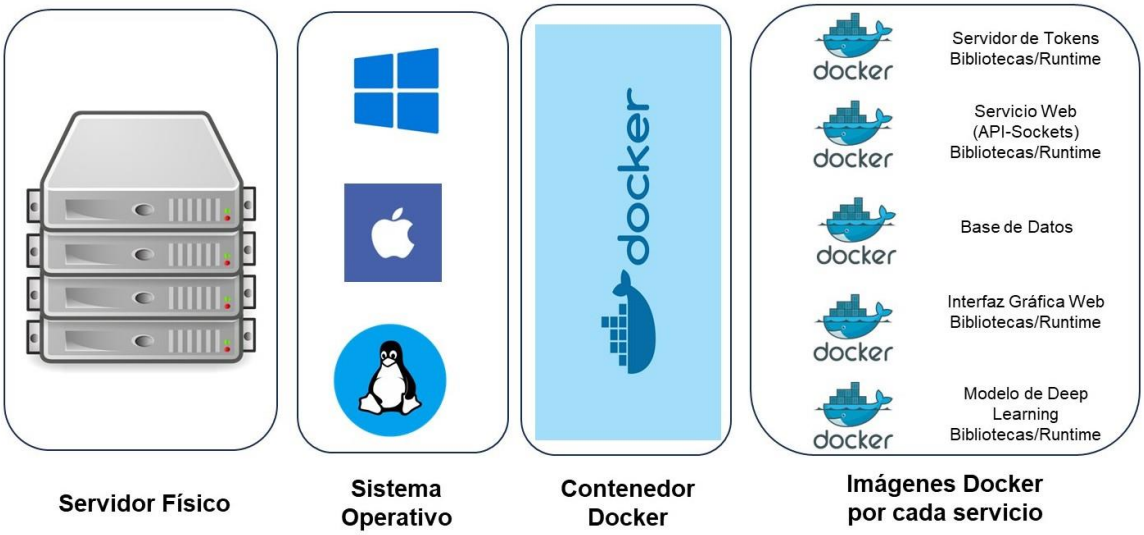


Figura 4-16: Interacción del sistema. En el servidor físico se instala el sistema operativo de preferencia, luego se monta el contenedor Docker como base para los servicios que se utilizarán. Por último, se monta una imagen Docker por cada servicio, las cuales interactuarán entre ellas.

Los microservicios se ejecutan dentro de un contenedor Docker mediante la creación de una imagen de Docker que contiene el código fuente del microservicio, sus dependencias y cualquier otro elemento necesario para su ejecución. La imagen de Docker se puede construir a partir de un archivo de configuración llamado *Dockerfile*, que especifica cómo se debe construir la imagen. Una vez que se ha creado la imagen, se puede utilizar para crear un contenedor Docker que ejecuta el microservicio. Ver Figura 4-16 para una descripción general.

4.5.3. Inferencia del modelo DL como servicio

Ofrecer inferencia de modelos como servicio es sencillo con una infraestructura moderna. Normalmente, los desarrolladores envuelven la función de inferencia del modelo detrás de una API que se puede llamar de forma remota, configuran ese servicio como un contenedor (por ejemplo, Docker, previamente analizado) e implementan el contenedor del servicio en máquinas virtuales o recursos de la nube.

En este trabajo un programa Python simple carga el modelo e implementa la función de inferencia del modelo como se analizó anteriormente. Luego se hace

uso de la librería *Flask* para aceptar solicitudes *Hypertext Transfer Protocol* HTTP en un puerto determinado y, para cada solicitud, ejecutar la función de inferencia del modelo y devolver el resultado como respuesta HTTP.

En esta configuración, el modelo se carga solo una vez cuando se inicia el proceso y luego puede atender solicitudes de inferencia con el mismo modelo. Se pueden iniciar múltiples procesos para compartir la carga.

Por supuesto, una API de este tipo también se puede diseñar para realizar múltiples predicciones en una sola llamada para ahorrar la sobrecarga de la red de llamadas individuales. Por ejemplo, un cliente podría enviar varias imágenes en una sola solicitud y recibir los objetos detectados para todas las imágenes en el resultado.

4.5.4. Interfaz web

En esta fase se crea una interfaz gráfica, la cual será usada por los especialistas en glaucoma, los cuales serán los usuarios finales y tendrán acceso a los microservicios antes descritos. A continuación se muestran un conjunto de interfaces que utilizan las tecnologías de *React* y *ViteJS* que facilitan la comunicación entre cliente y servidor sobre HTTP [157]. Las predicciones echas por el sistema automatizado pueden ser observadas en las siguientes figuras.

Figura 4-17: Interfaz de bienvenida del sistema automatizado.

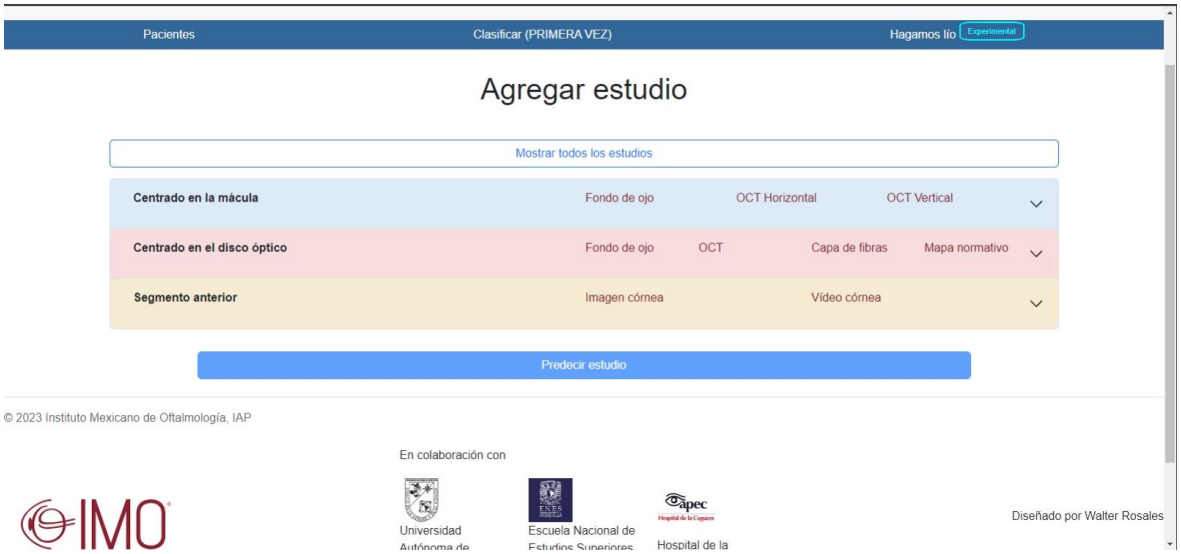


Figura 4-18: Selección del estudio.

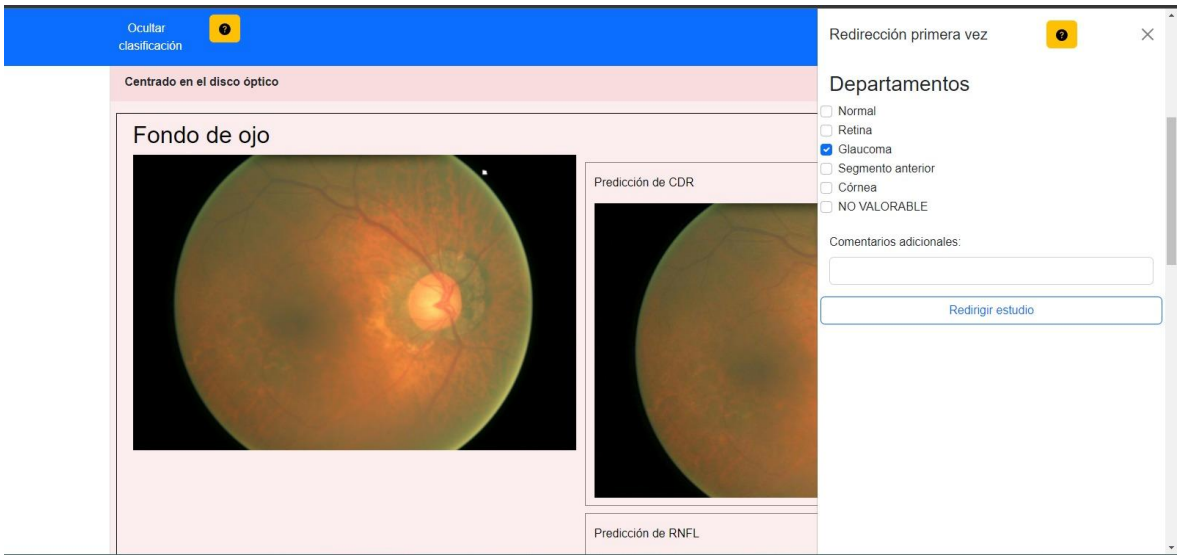


Figura 4-19: Selección del departamento para el cuál se realizará el estudio.

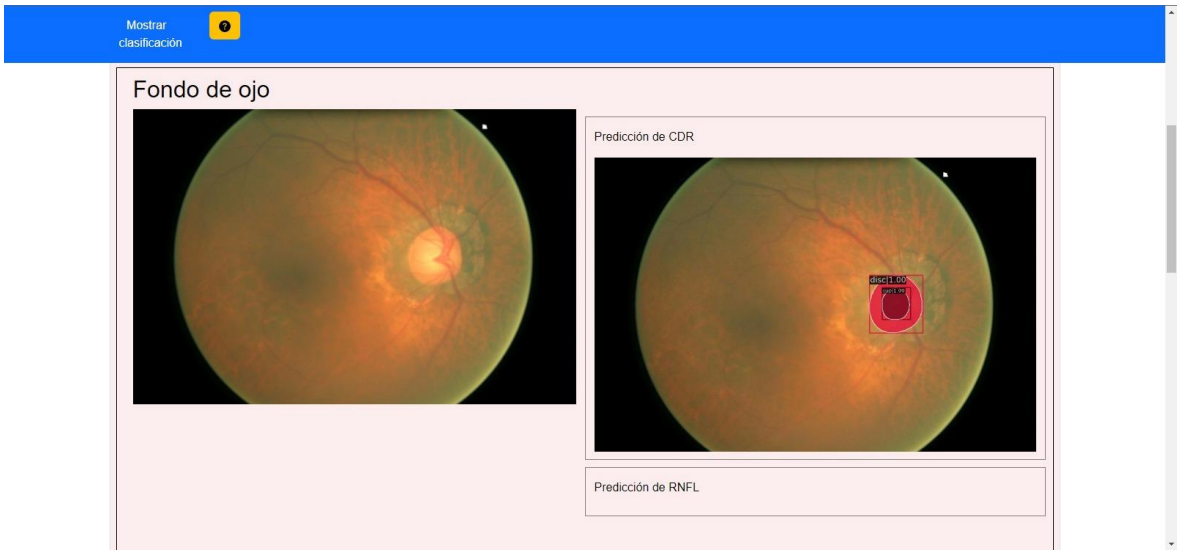


Figura 4-20: Imagen original (Izquierda) y segmentación del disco y la copa ópticas (Derecha).

Este es un proyecto en desarrollo, por lo que las imágenes previas son interfaces que pueden estar sugetas a cambios en el futuro.

Un último elemento a tener en cuenta fue elegir el entorno de producción adecuado. Esto implica decidir entre el alojamiento de servidores (en las instalaciones) y el alojamiento en la nube.

El alojamiento de servidores requiere una inversión inicial alta, pero ofrece más control, lo que lo convierte en la opción preferida para manejar datos confidenciales.

Por otro lado, el alojamiento en la nube puede ser más rentable y proporciona escalabilidad y flexibilidad. Como parte del acople del sistema automatizado al Instituto Mexicano de Oftalmología, se decidió integrar la producción a los servidores existentes en dicha institución. Otros factores fueron el tamaño y la naturaleza de los datos, la complejidad del modelo, consideraciones de costos y los requisitos de privacidad y seguridad de los datos.

4.6. Discusión

Se ha introducido un modelo de detección de objetos, con múltiples procesos de ajuste en el flujo de trabajo para detectar anomalías relacionadas con el glaucoma y la retinopatía diabética simultáneamente. La tarea se llevó a cabo en el conjunto de datos DDR, que aporta sus etiquetas y anotaciones para SE, EX, HE y MA. Además, puede añadir predicciones probabilísticas para las nuevas clases como la CO y el DO y la diferenciación entre las clases alfa y beta para la APP, características relacionadas con el glaucoma.

En el desarrollo de la metodología de esta investigación, compuesta por dos etapas, en la primera parte se estableció una comparación entre diferentes modelos de detección de objetos para la segmentación del DO y la CO en imágenes de fondo de ojo. Los modelos seleccionados fueron Mask R-CNN, Carafe, Cascade Mask R-CNN, GCNet, MS R-CNN, SOLO, and Point_Rend. Sus rendimientos fueron evaluados con las métricas AP, F1-Score y AUCPR.

La experimentación comenzó con un pequeño número de imágenes para observar el comportamiento del modelo con una cantidad reducida de imágenes, con el fin de obtener una buena segmentación tanto del DO como de la CO. Según [158], entre 150 y 500 imágenes existe un punto de inflexión en el que el rendimiento de los modelos comienza a aumentar significativamente. Inicialmente utilizamos 100 imágenes del conjunto de datos REFUGE para el entrenamiento, 30 para la validación y 30 para la prueba. Con esta proporción, los mejores rendimientos se asociaron con los modelos Cascade, Mask R-CNN y Point_Rend con un valor de 1.000 en AP y un umbral IoU de 0.50. Sin embargo, el resto de los

modelos tuvo un buen rendimiento, alrededor de 0.98, excepto SOLO, que resultó en un valor de 0.886; esto se debe a un mayor número de falsos positivos que son detecciones incorrectas de un objeto no existente o una detección mal colocada de un objeto existente. Según [159], los falsos positivos pueden estar relacionados con errores de clasificación, localización, clasificación más localización, duplicación y *background*. La ligera disminución en SOLO puede estar asociada con errores de localización cuando se detecta un objeto con un cuadro delimitador desalineado, lo que significa una superposición entre 0.1 y 0.5 IoU. Este problema puede abordarse utilizando muestras negativas en el proceso de entrenamiento, [160], [161]. El cálculo del F1-Score arrojó un valor perfecto de 1.

Con estos resultados previos, se procedió a anotar todo el conjunto de datos de REFUGE y se repitió la experimentación, donde MS R-CNN superó a los otros modelos, con una AP de 0.995 sin usar escala múltiple y 1.000 cuando se aplicó escala múltiple, siempre tomando el umbral IoU igual a 0.50. Esta técnica de aumento de datos mejoró ligeramente el rendimiento en todos los modelos excepto en el modelo SOLO, que disminuyó de 0.989 a 0.984, pero esto mejoró con el aumento de imágenes en comparación con la experimentación anterior. En cuanto al F1-Score, todos los modelos mantuvieron un valor perfecto de 1 excepto Cascade Mask R-CNN, que disminuyó a 0.997.

Experimentamos con estos resultados en el conjunto de datos G1020, ya mencionado anteriormente, obteniendo el mejor rendimiento el modelo Point_Rend con un valor de precisión media de 0,956. Los modelos restantes mantuvieron valores alrededor de 0,94 excepto el modelo SOLO, que disminuyó a 0,909. La disminución en el conjunto de datos puede estar relacionada con una heterogeneidad más significativa en sus imágenes; sin embargo, esto resultó en una ventaja al momento de transferir las predicciones a imágenes de otras fuentes que no se utilizaron en el proceso de entrenamiento. Por lo tanto, el conjunto de datos G1020 fue mejor para segmentar tanto el disco como la copa ópticos que REFUGE.

Un elemento clave utilizado en esta parte de la investigación fue la escala

múltiple, variando los valores entre 1333x640 a 1333x960. Este enfoque introdujo ligeras mejoras en las métricas cuando el número de imágenes aumentó, permitiendo la posibilidad de entrenar modelos con nuevas dimensiones. Sin embargo, este enfoque no trajo mejoras significativas con pocas imágenes, por lo que el investigador podría considerar su uso o no dependiendo de sus recursos.

Otro elemento que se ajustó fue la escala de anclaje (*anchor scale*). La configuración de las anclas está dada por *anchor_scales* y *anchor_ratios*, y originalmente se generaron cuatro en la RPN. Después de configurar *anchor_scales* con los valores siguientes, [4, 6, 8, 10, 12, 12, 14, 16] y *anchor_ratios* [0.5, 1.0, 1.5, 2.0], se generaron 28 anclas para cada nivel de extracción de características dentro de la FPN. Esta modificación nos permitió capturar una mayor diversidad en las formas del disco y las copas ópticas.

Una preocupación constante en el campo del ML es el desafío con el desequilibrio de datos. Si bien el desequilibrio de clases es una faceta de este problema que se discute con frecuencia, es simplemente una dimensión de un panorama de problemas más amplio. La segunda etapa de la metodología propuesta para esta investigación tiene como objetivo explorar diversas manifestaciones de este problema, incluidos los desequilibrios de *foreground-background*, los desequilibrios de *foreground-foreground*, los desequilibrios en la escala del nivel de caja, las disparidades en las distribuciones de Intersección sobre Unión (IoU), así como las inconsistencias en las distribuciones de pérdida de regresión.

En escenarios densos, como el enfoque de esta investigación, un número excesivo de muestras negativas puede resultar en una regresión inexacta de los cuadros delimitadores. Un desafío crucial es definir muestras positivas y negativas de alta calidad con precisión. El modo de cascada de umbrales múltiples de Cascade R-CNN proporciona una solución eficaz para mejorar la calidad de las regiones recomendadas y mitigar la influencia de ejemplos desafiantes en el resultado de detección final. Para abordar el desequilibrio en la distribución de IoU, se adoptó Cascade R-CNN como modelo primario, ya que, en lugar de

muestrearlos de nuevo, la utilización de cajas de la etapa anterior por parte de cada detector resalta la posibilidad de cambiar la distribución de IoU de sesgada a la izquierda a uniforme. Este modelo nos estregó resultados muy satisfactorios en la primera fase, que junto con sus características propias lo convierte en ideal para la segunda etapa.

Según los autores, la aplicación de NWD-RKA en Cascade R-CNN mejoró el AP en 7,1 puntos en el conjunto de pruebas AI-TOD-v2 [136]; esa fue la razón principal para utilizar esta técnica en esta investigación como referencia. Estos mecanismos mejoran el rendimiento sobre el estado actual arte en la detección de objetos diminutos en imágenes de retina, relacionados con la RD, al asignar más muestras positivas, mitigando los problemas de desequilibrio de escala y regresión, mientras se mantiene un alto rendimiento sobre objetos medios y grandes. El AP obtenido fue de 0,4295 con entropía cruzada como pérdida de la función de clasificación y un tamaño de imagen fijo de 1333 x 800.

El desequilibrio de *foreground-foreground* significa que algunos objetos están sobrerrepresentados o subrepresentados. Se utilizaron dos técnicas, OHEM y ASL, combinadas con aumento de MS, que van desde 1333 x 640 hasta 1333 x 960.

Al seleccionar automáticamente ejemplos desafiantes, OHEM simplifica el entrenamiento al eliminar la necesidad de diversas heurísticas e hiperparámetros comúnmente utilizados y mejora el rendimiento de prueba relacionado con mAP. Este enfoque mejora en comparación con el experimento base, como se ve en la Tabla 4-5.

Se observó la misma mejora con ASL y MS. ASL comprende dos mecanismos distintos y complementarios que funcionan de manera diferente cuando se aplican a muestras positivas y negativas. Al realizar un análisis de probabilidad de detección de la red, hemos demostrado la eficacia de ASL para lograr un equilibrio armonioso entre muestras negativas y positivas.

WDIoUNMS demuestra una alta eficiencia. Además, la incorporación de factores geométricos mejora el AP y este enfoque mejora constantemente el rendimiento al

entrenar modelos profundos para la regresión de cuadros delimitadores. Al agregar esta técnica, con doce épocas y un tamaño de lote de dos, los resultados fueron 0,446 de AP para CE+OHEM+MS+WDIoUNMS y 0,436 para ASL+MS+WDIoUNMS.

Cuando se introdujeron más épocas (cincuenta) y un tamaño de lote más grande (ocho), el resultado informado en esta investigación fue para CE+OHEM+MS+WDIoUNMS y ASL+MS+WDIoUNMS 0.451 y 0.460, respectivamente. Además, FPN fue el esquema base en el componente de *Neck*; se realizaron algunas experimentaciones con PAFPN, pero no obtuvieron mejores resultados. FPN ayuda fundamentalmente con el problema de desequilibrio de escala a nivel de caja.

Para el análisis de la evaluación, el marco TIDE resalta en detalle la principal contribución al error, mostrando experimentos basados en ASL que mejoran el manejo de los errores relacionados con la clasificación y la localización.

Establecer una comparación justa con el estado del arte fue difícil. Los trabajos citados en esta investigación muestran principalmente resultados relacionados con el conjunto de datos de DDR con sus anotaciones originales. Aquí se agregaron características relacionadas con la enfermedad del glaucoma, como la copa y el disco ópticos, y la atrofia peripapilar (alfa/beta), por lo que se informó que el mAP tenía más clase que promediar. Sin embargo, se puede proporcionar cierta aproximación extrayendo información de la matriz de confusión.

Santos et al. [103] muestran su matriz de confusión, para sus experimentos con el optimizador Adam, sobre el conjunto de datos de validación, información que la red neuronal utiliza para ajustar el modelo, y por eso, no es ideal. Según eso, reportan valores de precisión para SE, EX, HE y MA iguales a 0.35, 0.33, 0.32 y 0.18, respectivamente. En este trabajo, la misma información se puede extraer de la matriz de confusión, pero el modelo no ve información sobre el conjunto de prueba. Los resultados para SE, EX, HE y MA fueron 0.35, 0.41, 0.18 y 0.21, respectivamente, para ASL+MS+WDIoUNMS, el mejor rendimiento en esta

investigación. Dado que los autores citados informan un mAP para el conjunto de validación de 0.2630 y para el conjunto de prueba de 0.1540, y las métricas reportadas en esta investigación son cercanas y, en algunos casos, superan los valores de su conjunto de validación, se puede concluir que este trabajo logra resultados comparables o incluso mejores.

La metodología presentada introduce un nuevo enfoque en el que se emplea un modelo para la detección de anomalías. Dada su naturaleza como modelo de detección de objetos, se elimina la necesidad de un mapa saliente para acentuar las lesiones. Esto se debe al objetivo fundamental de esta categoría de modelos, que gira en torno a encapsular la lesión identificada dentro de un cuadro delimitador.

Finalmente, la integración en un entorno de producción permitió que médicos y especialistas del área evaluaran los resultados obtenidos, demostrando la importancia de implementar modelos de DL.

Si bien algunos modelos pueden interpretarse sin implementación, muchos requieren configuraciones específicas para funcionar de manera óptima, como ser parte de una aplicación o un pipeline integrado. A menudo, esto se puede lograr colocando el modelo a ser consumido por una API, lo que le permite interactuar con otros componentes de desarrollo de software. Este fue el enfoque implementado a través de una arquitectura de microservicios y el uso de Docker como contenedores, para la estandarización de los recursos.

La implementación de modelos de DL permite a las empresas e instituciones aprovechar el poder de la IA para impulsar los resultados al:

- Mejorar la eficiencia mediante la automatización de tareas repetitivas, lo que genera ahorros de costos significativos.
- Mejorar la toma de decisiones utilizando las predicciones y los conocimientos precisos del modelo.
- Descubrir patrones y tendencias ocultos en los datos y proporcionar

información valiosa que de otro modo podría permanecer sin descubrir.

5. Conclusiones

Arribando al final de esta investigación retomamos su inicio, donde nos planteamos la incógnita científica de si la detección de múltiples biomarcadores de daño en imágenes de retina, a través de técnicas de DL, generará patrones y predicciones de enfermedades crónicas en la retina que deriven en clasificaciones. Esta incógnita generó el objetivo general de desarrollar un marco integral que permitiera la detección de lesiones en la retina, en esta investigación asociadas al glaucoma y la retinopatía diabética.

Para darle cumplimiento a los objetivos planteados, se inició con la segmentación del disco y copa ópticas, a través de la comparación de múltiples modelos de detección de objetos. Esto crea precedente en el estado del arte, ya que se sale del enfoque tradicional basado en la segmentación a partir de estructuras de tipo encoder-decoder. Las bases de datos elegidas fueron REFUGE y G1020, aplicando un efectivo mecanismo de anotación sobre la primera. Se contestó a la pregunta de cuantas imágenes iniciales son necesarias para obtener buenos resultados, empleando 100 y obteniendo resultados satisfactorios en la segmentación de las estructuras antes mencionadas; sin embargo, utilizar todo el conjunto de datos trajo una ligera mejoría.

Otro problema común abordado en este trabajo fue la reducción de imágenes. La mayoría de los trabajos de vanguardia reducen significativamente las imágenes de entrada o introducen un nuevo flujo de trabajo en el proceso de segmentación al recortar el área de interés. Este trabajo, con aumento de datos basado en múltiples escalas, demuestra que mejora los resultados, manteniendo imágenes de entrada de alta resolución y evitando reducciones significativas. El ajuste de cajas de ancla para propuestas regionales también mejora la ubicación del objetivo. El optimizador AdamW y la estrategia de recocido sinusoidal en el programa de aprendizaje también mejoraron los resultados. De este proceso se fue capaz de identificar al modelo Cascade Mask R-CNN como candidato a crear un marco de

trabajo integral para la detección de lesiones en la retina.

Precisamente el modelo Cascade R-CNN fue elegido base para darle cumplimiento al objetivo específico número dos. Este modelo fue elegido por su capacidad para aumentar el umbral de IoU, obteniendo muestras de mayor calidad en cada nivel. Además de eso, Cascade R-CNN permite la detección de múltiples clases simultáneamente, una ventaja sobre otros trabajos de vanguardia que aplican múltiples modelos, uno por lesión, lo que complica el marco de detección general.

Junto al modelo en cascada; se aplicaron nuevos métodos para manejar los problemas de desequilibrio de datos. La normalización de la distancia de Wasserstein con su esquema de asignación basado en rangos demuestra una alta efectividad con objetos diminutos. El mejor resultado general proviene de la pérdida asimétrica para la selección de clases difíciles y un grupo ponderado con distancia IoU en la técnica de supresión no máxima de posprocesamiento. La precisión promedio media fue de 0.461.

Otra contribución es la efectividad del etiquetado suave para ayudar con la necesidad de conjuntos de datos etiquetados, una tarea que consume mucho tiempo para la mayoría de los especialistas. Aquí se aplicó a la atrofia peripapilar y, según nuestro conocimiento, es la primera vez que se detecta y separa por sus clases, alfa, beta, estando la última relacionada con la progresión del glaucoma.

Para darle cumplimiento a los objetivos específicos tres y cuatro, se desarrolló una plataforma de software que permite la correcta visualización de diferentes lesiones y la continua evaluación de los resultados por especialistas del área. Además, se estableció con comparación con el estado del arte en la medida de lo posible por la novedad de esta investigación en detectar múltiples lesiones en imágenes de retina, asociadas a diferentes enfermades. Por ejemplo, el trabajo de [50] mostró un mayor valor de F1-Score, pero ellos realizaron la tarea de segmentar el disco y la copa ópticas por separado, lo que es menos retador para los modelos, mientras que aquí se hizo de forma conjunta. Por otro lado, en la detección de las

lesiones, la comparación fue más compleja debido a la novedad de esta investigación; sin embargo, con el apoyo de matrices de confusión, fue posible separar las detecciones por clases y evidenciar las mejoras que introdujo la metodología propuesta en esta investigación.

Por tanto, se logró validar la hipótesis planteada en el trabajo, al ser capaz de detectar lesiones en la retina, asociadas a múltiples enfermedades, a través de modelos de DL, mientras mantiene resultados equiparables o superiores a los de investigaciones previas. Se da respuesta a la incógnita científica además de que sienta bases para futuros trabajos donde colaboren equipos multidisciplinares.

6. Recomendaciones y trabajos futuros

Algunas limitaciones fueron identificadas a lo largo de la investigación. Debido a la complejidad de evaluar las segmentaciones, es difícil establecer comparaciones y seleccionar las mejores arquitecturas, ya que los investigadores utilizan diferentes métricas, además de diferentes fuentes de datos, que incluso bajo las mismas condiciones de entrenamiento y prueba, hacen que sea imposible comparar dos trabajos si no se utiliza el mismo conjunto de datos. Otra limitación fue la imposibilidad de entrenar todos los modelos, con las diferentes *backbones* que se pueden utilizar, debido a la capacidad computacional y la tarea que llevaría mucho tiempo probar todas las variantes.

Otra limitación en este estudio viene con un desequilibrio de clases significativo. El número de instancias para exudados duros fue de 11136, y la clase con menor representación fue la atrofia alfa con 126 instancias. Las disparidades en las dimensiones de los objetos presentan un desafío adicional en escenarios donde existe una disparidad sustancial entre el recuento de píxeles atribuido a la clase de DO y el de la clase de MA.

Al igual que los trabajos académicos a los que se alude en esta investigación, la identificación de HE y MA sigue siendo una preocupación importante. La razón es el tamaño, porque son características diminutas en la mayoría de los casos, y a menudo se confunden con el fondo debido a su color natural.

El proceso de anotación por parte de especialistas está lleno de variabilidad que puede estar relacionada tanto con factores humanos como con la calidad de la imagen. La disponibilidad de anotaciones para nuevas enfermedades es escasa en la mayoría de los conjuntos de datos. La mayoría de las veces, estos conjuntos de datos están especializados, por lo que es importante evitar no anotar objetos que ya están presentes en el conjunto de datos para no perjudicar el proceso de aprendizaje.

En futuros trabajos, se pretende manejar el desequilibrio de clases. Dado que el submuestreo puede perjudicar el proceso de aprendizaje al introducir falsos negativos, el sobre muestreo de clases minoritarias podría ser un enfoque; las técnicas generativas como las GAN son un ejemplo. El uso de parches del conjunto de datos es otra opción para evitar que los objetos pequeños desaparezcan durante el proceso de reducción de escala, pero es necesario vigilar la anotación de cajas delimitadoras para los objetos más grandes.

Los desequilibrios de datos aún deben resolverse con técnicas de DL. El estudio de nuevas arquitecturas será necesario. Además, esta investigación crea oportunidades para desarrollar sistemas para clínicas en comunidades con un bajo costo computacional, ya que un modelo puede detectar múltiples hallazgos a la vez, y esto podría hacer una contribución significativa a la reducción de costos para el cribado de fondo de ojo en regiones donde los oftalmólogos son escasos, lo que lleva a la detección temprana de numerosas enfermedades potencialmente amenazantes para la vista.

Finalmente, se hace necesario crear sistemas inteligentes que cuenten con la capacidad de explicar sus decisiones, por lo que se explorarán los modelos causales como mecanismo de dicha explicación. Los modelos causales son modelos matemáticos que representan las relaciones causales dentro de un sistema; son herramientas diseñadas para simplificar el entendimiento de sistemas complejos, facilitando su comprensión. Pueden enseñarnos mucho sobre la epistemología de la causalidad, y sobre la relación entre causalidad y probabilidad [162], [163].

7. Referencias bibliográficas

[1] World Health Organisation, *World report on vision*, vol. 214, no. 14. 2019.

[2] A. Avellaneda, M. Izquierdo, J. Torrent-Farnell, and J. R. Ramón, “Enfermedades raras: Enfermedades crónicas que requieren un nuevo enfoque sociosanitario,” *An. Sist. Sanit. Navar.*, vol. 30, no. 2, pp. 177–190, 2007,doi:10.4321/s1137-66272007000300002.

[3] “análisis | Definición | Diccionario de la lengua española | RAE - ASALE.” https://dle.rae.es/análisis (accessed Sep. 28, 2020).

[4] S. Qian, “What Is Detection?,” *Detection*, vol. 02, no. 02, pp. 7–9, 2014,doi:10.4236/detection.2014.22002.

[5] M. D. Abramoff, M. K. Garvin, and M. Sonka, “Retinal Imaging and Image Analysis,” *IEEE Rev. Biomed. Eng.*, vol. 3, pp. 169–208, 2010, [Online]. Available: http://ieeexplore.ieee.org/document/5660089/, doi:10.1109/RBME.2010.2084567.

[6] D. S. W. Ting and T. Y. Wong, “Eyeing cardiovascular risk factors,” *Nat. Biomed. Eng.*, vol. 2, no. 3, pp. 140–141, 2018, [Online]. Available: http://dx.doi.org/10.1038/s41551-018-0210-5, doi:10.1038/s41551-018-0210-5.

[7] C. Outline, *Deep learning - 10 rnn*. 2020.doi:10.1016/B978-0-12-804291-5.00010-6.

[8] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015,doi:10.1038/nature14539.

[9] “MAILOR.” https://mailor.com.mx/ (accessed Sep. 29, 2020).

[10] I. Torres Courchoud and J. I. Pérez Calvo, “Biomarcadores y práctica clínica,” *An. Sist. Sanit. Navar.*, vol. 39, no. 1, pp. 5–8, 2016,doi:10.4321/S1137-6627/2016000100001.

[11] "Webvision – The Organization of the Retina and Visual System." https://webvision.med.utah.edu/ (accessed Oct. 05, 2023).

[12] "IDF Diabetes Atlas 2021 | IDF Diabetes Atlas." https://diabetesatlas.org/atlas/tenth-edition/ (accessed Oct. 06, 2023).

[13] R. R. A. Bourne *et al.*, "Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: The Right to Sight: An analysis for the Global Burden of Disease Study," *Lancet Glob. Heal.*, vol. 9, no. 2, pp. e144–e160, Feb. 2021, Accessed: Aug. 12, 2023. [Online]. Available: http://www.thelancet.com/article/S2214109X20304897/fulltext, doi:10.1016/S2214-109X(20)30489-7.

[14] "Retinopatia Diabetica: Un Reto Para la Salud Visual en Mexico - The International Agency for the Prevention of Blindness." https://www.iapb.org/news/retinopatia-diabetica-un-reto-para-la-salud-visual-en-mexico/ (accessed Oct. 06, 2023).

[15] J. A. Giaconi, S. K., A. L. Coleman, J. Caprioli, and K. Nouri-Mahdavi, Eds., *Pearls of glaucoma management*, 2nd ed. Berlin, Germany: Springer, 2016.

[16] "123. Se estima que en México 1.5 millones de personas padecen glaucoma: Secretaría de Salud | Secretaría de Salud | Gobierno | gob.mx." https://www.gob.mx/salud/prensa/123-se-estima-que-en-mexico-1-5-millones-de-personas-padecen-glaucoma-secretaria-de-salud (accessed Oct. 06, 2023).

[17] H. Quigley and A. T. Broman, "The number of people with glaucoma worldwide in 2010 and 2020," *Br. J. Ophthalmol.*, vol. 90, no. 3, pp. 262–267, Mar. 2006, Accessed: Jun. 21, 2022. [Online]. Available: https://bjo.bmj.com/content/90/3/262, doi:10.1136/BJO.2005.081224.

[18] E. N. Oftalmología, "Prevención de las patologías oculares en oftalmología," pp. 49–50, 2014.

[19] A. Grzybowski *et al.*, “Artificial intelligence for diabetic retinopathy screening: a review,” *Eye*, vol. 34, no. 3, p. 451, Mar. 2020, Accessed: Oct. 13, 2023. [Online]. Available: /pmc/articles/PMC7055592/, doi:10.1038/S41433-019-0566-0.

[20] “Eyenuk Secures the First European Union MDR Certification.” https://www.globenewswire.com/en/news-release/2023/01/31/2598236/0/en/Eyenuk-Secures-the-First-European-Union-MDR-Certification-for-Autonomous-AI-Detection-of-Diabetic-Retinopathy-Age-Related-Macular-Degeneration-and-Glaucoma.html (accessed Oct. 13, 2023).

[21] T. Narendra, A. Sankaran, D. Vijaykeerthy, and S. Mani, “Explaining Deep Learning Models using Causal Inference,” 2018, [Online]. Available: http://arxiv.org/abs/1811.04376,

[22] J. J. Kanski and B. Bowling, *Kanski’s Clinical Ophthalmology E-Book: A Systematic Approach*. Elsevier Health Sciences, 2015. [Online]. Available: https://books.google.es/books?id=D9GfBwAAQBAJ,

[23] P. Riordan-Eva and J. J. Augsburger, “Vaughan & Asbury’s General Ophthalmology, 19e,” 2018.

[24] Y. X. Wang, S. Panda-Jonas, and J. B. Jonas, “Optic nerve head anatomy in myopia and glaucoma, including parapapillary zones alpha, beta, gamma and delta: Histology and clinical features,” *Prog. Retin. Eye Res.*, vol. 83, Jul. 2021, Accessed: Oct. 30, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33309588/, doi:10.1016/J.PRETEYERES.2020.100933.

[25] E. Bernal-Catalán, E. De la Cruz-Gámez, J. A. Montero-Valverde, R. H. Reyna, and J. L. Hernandez-Hernández, “Detection of Exudates and Microaneurysms in the Retina by Segmentation in Fundus Images,” *Rev. Mex. Ing. Biomed.*, vol. 42, no. 2, pp. 67–77, May 2021, Accessed: Oct. 30, 2023.

[Online]. Available: https://www.rmib.mx/index.php/rmib/article/view/1136, doi:10.17488/RMIB.42.2.6.

[26] M. Arbib, D. Ballard, J. Bower, and G. Orban, *Neural networks : algorithms, applications, and programming techniques /*, vol. 7, no. 1. Addison-Wesley, 1991. Accessed: Jan. 06, 2024. [Online]. Available: http://www.amazon.com/dp/0201513765,

[27] G. Litjens *et al.*, “A survey on deep learning in medical image analysis,” *Med. Image Anal.*, vol. 42, no. December 2012, pp. 60–88, Dec. 2017, [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1361841517301135, doi:10.1016/j.media.2017.07.005.

[28] B. M. Brio and A. S. Molina, *Redes Neuronales y Sistemas Borrosos. 3a Edición*. RA-MA S.A. Editorial y Publicaciones, 2006. [Online]. Available: https://books.google.com.mx/books?id=KwbjGAAACAAJ,

[29] C. C. Aggarwal, *Neural Networks and Deep Learning*. 2018.doi:10.1007/978-3-319-94463-0.

[30] I. Goodfellow, Y. Bengio, and A. Courville, “Deep Learning,” *Nat. Methods*, vol. 13, no. 1, pp. 35–35, 2015, [Online]. Available: http://www.nature.com/doifinder/10.1038/nature14539%5Cnhttp://www.nature.com/doifinder/10.1038/nmeth.3707, doi:10.1038/nmeth.3707.

[31] “A Gentle Introduction to the Rectified Linear Unit (ReLU).” https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/ (accessed May 20, 2020).

[32] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *32nd Int. Conf. Mach. Learn. ICML 2015*, vol. 1, pp. 448–456, 2015.

[33] G. W. Lindsay, “Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future,” *J. Cogn. Neurosci.*, pp. 1–15,

2020,doi:10.1162/jocn_a_01544.

[34] “Convolutional Neural Networks (CNN): Summary - Blogs SuperDataScience - Big Data | Analytics Careers | Mentors | Success.” https://www.superdatascience.com/blogs/convolutional-neural-networks-cnn-summary (accessed Sep. 29, 2020).

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” *Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, Accessed: Oct. 31, 2023. [Online]. Available: http://code.google.com/p/cuda-convnet/,

[36] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, “ImageNet: A large-scale hierarchical image database,” pp. 248–255, Mar. 2010,doi:10.1109/CVPR.2009.5206848.

[37] F. Zhuang *et al.*, “A Comprehensive Survey on Transfer Learning,” *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021,doi:10.1109/JPROC.2020.3004555.

[38] “CS231n Convolutional Neural Networks for Visual Recognition.” https://cs231n.github.io/transfer-learning/ (accessed Oct. 31, 2023).

[39] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” Nov. 2013, [Online]. Available: http://arxiv.org/abs/1311.2524,

[40] R. Girshick, “Fast R-CNN,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, vol. 2015 Inter, pp. 1440–1448. [Online]. Available: http://ieeexplore.ieee.org/document/7410526/, doi:10.1109/ICCV.2015.169.

[41] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” Jun. 2015, Accessed: Jun. 18, 2022. [Online]. Available: http://arxiv.org/abs/1506.01497,

[42] K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask R-CNN,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, vol. 2017-Octob, pp. 2980–2988. [Online]. Available: http://ieeexplore.ieee.org/document/8237584/, doi:10.1109/ICCV.2017.322.

[43] “Python Software Foundation,” 2017. https://www.python.org/psf/ (accessed Jun. 04, 2020).

[44] “Lista de licencias con comentarios - Proyecto GNU - Free Software Foundation.” http://www.gnu.org/licenses/license-list.html#PythonOld (accessed Jun. 04, 2020).

[45] “PyTorch.” https://pytorch.org/ (accessed Jun. 04, 2020).

[46] “Anaconda | The World’s Most Popular Data Science Platform.” https://www.anaconda.com/ (accessed Jun. 04, 2020).

[47] “Spyder Website.” https://www.spyder-ide.org/ (accessed Jun. 04, 2020).

[48] K. Chen *et al.*, “MMDetection: Open MMLab Detection Toolbox and Benchmark,” Jun. 2019, [Online]. Available: http://arxiv.org/abs/1906.07155,

[49] J. I. Orlando *et al.*, “REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs,” *Med. Image Anal.*, vol. 59, p. 101570, Jan. 2020, [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1361841519301100, doi:10.1016/j.media.2019.101570.

[50] M. N. Bajwa, G. A. P. Singh, W. Neumeier, M. I. Malik, A. Dengel, and S. Ahmed, “G1020: A Benchmark Retinal Fundus Image Dataset for Computer-Aided Glaucoma Detection,” in *2020 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2020, pp. 1–7. [Online]. Available: https://ieeexplore.ieee.org/document/9207664/, doi:10.1109/IJCNN48605.2020.9207664.

[51] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, “LabelMe: A

Database and Web-Based Tool for Image Annotation," *Int. J. Comput. Vis.*, vol. 77, no. 1–3, pp. 157–173, May 2008, [Online]. Available: http://link.springer.com/10.1007/s11263-007-0090-8, doi:10.1007/s11263-007-0090-8.

[52] Zhuo Zhang *et al.*, "ORIGA^{-light}: An online retinal fundus image database for glaucoma analysis and research," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, Aug. 2010, vol. 2010, pp. 3065–3068. [Online]. Available: http://ieeexplore.ieee.org/document/5626137/, doi:10.1109/IEMBS.2010.5626137.

[53] "Roboflow: Give your software the power to see objects in images and video." https://roboflow.com/ (accessed Jun. 14, 2023).

[54] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, and H. Kang, "Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening," *Inf. Sci. (Ny).*, vol. 501, pp. 511–522, 2019, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025519305377, doi:https://doi.org/10.1016/j.ins.2019.06.011.

[55] J. B. Jonas, P. Martus, F. K. Horn, A. Jünemann, M. Korth, and W. M. Budde, "Predictive factors of the optic nerve head for development or progression of glaucomatous visual field loss," *Invest. Ophthalmol. Vis. Sci.*, vol. 45, no. 8, pp. 2613–2618, 2004, Accessed: Jun. 03, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/15277484/, doi:10.1167/IOVS.03-1274.

[56] A. Septiarini and A. Harjoko, "AUTOMATIC GLAUCOMA DETECTION BASED ON THE TYPE OF FEATURES USED: A REVIEW," *J. Theor. Appl. Inf. Technol.*, vol. 28, no. 3, 2015, Accessed: Jun. 03, 2023. [Online]. Available: www.jatit.org,

[57] H. N. Veena, A. Muruganandham, and T. S. Kumaran, "A Review on the optic disc and optic cup segmentation and classification approaches over retinal

fundus images for detection of glaucoma," *SN Appl. Sci.*, vol. 2, no. 9, p. 1476, 2020, [Online]. Available: https://doi.org/10.1007/s42452-020-03221-z, doi:10.1007/s42452-020-03221-z.

[58] M. Alawad *et al.*, "Machine Learning and Deep Learning Techniques for Optic Disc and Cup Segmentation – A Review," *Clin. Ophthalmol.*, vol. Volume 16, pp. 747–764, Mar. 2022, [Online]. Available: https://www.dovepress.com/machine-learning-and-deep-learning-techniques-for-optic-disc-and-cup-s-peer-reviewed-fulltext-article-OPTH, doi:10.2147/OPTH.S348479.

[59] X. Sun, Y. Xu, W. Zhao, T. You, and J. Liu, "Optic Disc Segmentation from Retinal Fundus Images via Deep Object Detection Networks," *Conf. Proc. ... Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf.*, vol. 2018, pp. 5954–5957, 2018,doi:10.1109/EMBC.2018.8513592.

[60] A. Chakravarty and J. Sivaswamy, "A Deep Learning based Joint Segmentation and Classification Framework for Glaucoma Assesment in Retinal Color Fundus Images," Jul. 2018, Accessed: Nov. 11, 2022. [Online]. Available: https://arxiv.org/abs/1808.01355v1, doi:10.48550/arxiv.1808.01355.

[61] B. Al-Bander, B. M. Williams, W. Al-Nuaimy, M. A. Al-Taee, H. Pratt, and Y. Zheng, "Dense Fully Convolutional Segmentation of the Optic Disc and Cup in Colour Fundus for Glaucoma Diagnosis," *Symmetry 2018, Vol. 10, Page 87*, vol. 10, no. 4, p. 87, Mar. 2018, Accessed: Nov. 11, 2022. [Online]. Available: https://www.mdpi.com/2073-8994/10/4/87/htm, doi:10.3390/SYM10040087.

[62] Z. Gu *et al.*, "CE-Net: Context Encoder Network for 2D Medical Image Segmentation," *IEEE Trans. Med. Imaging*, vol. 38, no. 10, pp. 2281–2292, 2019,doi:10.1109/TMI.2019.2903562.

[63] H. Fu *et al.*, "Disc-Aware Ensemble Network for Glaucoma Screening From

Fundus Image," *IEEE Trans. Med. Imaging*, vol. 37, no. 11, pp. 2493–2501, 2018,doi:10.1109/TMI.2018.2837012.

[64] V. K. Singh *et al.*, "Retinal Optic Disc Segmentation using Conditional Generative Adversarial Network," *Front. Artif. Intell. Appl.*, vol. 308, pp. 373–380, Jun. 2018, Accessed: Nov. 11, 2022. [Online]. Available: https://arxiv.org/abs/1806.03905v1, doi:10.48550/arxiv.1806.03905.

[65] S. Sreng, N. Maneerat, K. Hamamoto, and K. Y. Win, "Deep learning for optic disc segmentation and glaucoma diagnosis on retinal images," *Appl. Sci.*, vol. 10, no. 14, 2020,doi:10.3390/app10144916.

[66] S. Wang, L. Yu, X. Yang, C.-W. Fu, and P.-A. Heng, "Patch-Based Output Space Adversarial Learning for Joint Optic Disc and Cup Segmentation," *IEEE Trans. Med. Imaging*, vol. 38, no. 11, pp. 2485–2495, 2019,doi:10.1109/TMI.2019.2899910.

[67] J. Son, S. J. Park, and K.-H. Jung, "Towards Accurate Segmentation of Retinal Vessels and the Optic Disc in Fundoscopic Images with Generative Adversarial Networks," *J. Digit. Imaging*, vol. 32, no. 3, pp. 499–512, 2019, [Online]. Available: https://doi.org/10.1007/s10278-018-0126-3, doi:10.1007/s10278-018-0126-3.

[68] B. Liu, D. Pan, and H. Song, "Joint optic disc and cup segmentation based on densely connected depthwise separable convolution deep network," *BMC Med. Imaging*, vol. 21, no. 1, pp. 1–12, 2021, [Online]. Available: https://doi.org/10.1186/s12880-020-00528-6, doi:10.1186/s12880-020-00528-6.

[69] Z. Tian, Y. Zheng, X. Li, S. Du, and X. Xu, "Graph convolutional network based optic disc and cup segmentation on fundus images," *Biomed. Opt. Express*, vol. 11, no. 6, pp. 3043–3057, Jun. 2020, [Online]. Available: https://opg.optica.org/boe/abstract.cfm?URI=boe-11-6-3043, doi:10.1364/BOE.390056.

[70] Y. Zheng, X. Zhang, X. Xu, Z. Tian, and S. Du, “Deep level set method for optic disc and cup segmentation on fundus images,” *Biomed. Opt. Express*, vol. 12, no. 11, pp. 6969–6983, Nov. 2021, [Online]. Available: https://opg.optica.org/boe/abstract.cfm?URI=boe-12-11-6969, doi:10.1364/BOE.439713.

[71] J. Zhang, Y. Zheng, W. Hou, and W. Jiao, “Leveraging non-expert crowdsourcing to segment the optic cup and disc of multicolor fundus images,” *Biomed. Opt. Express*, vol. 13, no. 7, pp. 3967–3982, Jul. 2022, [Online]. Available: https://opg.optica.org/boe/abstract.cfm?URI=boe-13-7-3967, doi:10.1364/BOE.461775.

[72] R. Yang and Y. Yu, “Artificial Convolutional Neural Network in Object Detection and Semantic Segmentation for Medical Imaging Analysis,” *Front. Oncol.*, vol. 11, p. 573, Mar. 2021,doi:10.3389/FONC.2021.638182/BIBTEX.

[73] M. Orouskhani, N. Firoozeh, S. Xia, M. Mossa-Basha, and C. Zhu, “nnDetection for Intracranial Aneurysms Detection and Localization,” May 2023, Accessed: Jun. 08, 2023. [Online]. Available: https://arxiv.org/abs/2305.13398v1,

[74] F. Hardalaç *et al.*, “Fracture Detection in Wrist X-ray Images Using Deep Learning-Based Object Detection Models,” *Sensors (Basel).*, vol. 22, no. 3, Feb. 2022, Accessed: Jun. 08, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/35162030/, doi:10.3390/S22031285.

[75] T. Hirasawa *et al.*, “Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images,” *Gastric Cancer*, vol. 21, no. 4, pp. 653–660, Jul. 2018, Accessed: Jun. 08, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29335825/, doi:10.1007/S10120-018-0793-2.

[76] L. Wu *et al.*, “A deep neural network improves endoscopic detection of early gastric cancer without blind spots,” *Endoscopy*, vol. 51, no. 6, pp. 522–531,

2019, Accessed: Sep. 15, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/30861533/, doi:10.1055/A-0855-3532.

[77] S. Ali *et al.*, “Endoscopy artifact detection (EAD 2019) challenge dataset,” May 2019, Accessed: Jun. 08, 2023. [Online]. Available: http://arxiv.org/abs/1905.03209, doi:10.17632/C7FJBXCGJ9.1.

[78] D. Jha *et al.*, “Real-Time Polyp Detection, Localization and Segmentation in Colonoscopy Using Deep Learning,” *Ieee Access*, vol. 9, p. 40496, 2021, Accessed: Jun. 08, 2023. [Online]. Available: /pmc/articles/PMC7968127/, doi:10.1109/ACCESS.2021.3063716.

[79] Y. Liu *et al.*, “Detecting Cancer Metastases on Gigapixel Pathology Images,” Mar. 2017, Accessed: Jun. 08, 2023. [Online]. Available: https://arxiv.org/abs/1703.02442v2,

[80] N. Tomita, B. Abdollahi, J. Wei, B. Ren, A. Suriawinata, and S. Hassanpour, “Attention-Based Deep Neural Networks for Detection of Cancerous and Precancerous Esophagus Tissue on Histopathological Slides,” *JAMA Netw. open*, vol. 2, no. 11, Nov. 2019, Accessed: Jun. 08, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31693124/, doi:10.1001/JAMANETWORKOPEN.2019.14645.

[81] K. Yan *et al.*, “MULAN: Multitask Universal Lesion Analysis Network for Joint Lesion Detection, Tagging, and Segmentation,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11769 LNCS, pp. 194–202, 2019, Accessed: Jun. 08, 2023. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-32226-7_22, doi:10.1007/978-3-030-32226-7_22/COVER.

[82] M. Zlocha, Q. Dou, and B. Glocker, “Improving RetinaNet for CT Lesion Detection with Dense Masks from Weak RECIST Labels,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11769 LNCS, pp. 402–410, 2019, Accessed: Jun. 08,

2023. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-32226-7_45, doi:10.1007/978-3-030-32226-7_45/COVER.

[83] Y. Shou, T. Meng, W. Ai, C. Xie, H. Liu, and Y. Wang, “Object Detection in Medical Images Based on Hierarchical Transformer and Mask Mechanism,” *Comput. Intell. Neurosci.*, vol. 2022, 2022, Accessed: Jun. 08, 2023. [Online]. Available: /pmc/articles/PMC9371842/, doi:10.1155/2022/5863782.

[84] Q. T. Huynh *et al.*, “Automatic Acne Object Detection and Acne Severity Grading Using Smartphone Images and Artificial Intelligence,” *Diagnostics 2022, Vol. 12, Page 1879*, vol. 12, no. 8, p. 1879, Aug. 2022, Accessed: Jun. 08, 2023. [Online]. Available: https://www.mdpi.com/2075-4418/12/8/1879/htm, doi:10.3390/DIAGNOSTICS12081879.

[85] A. K. Tyagi *et al.*, “DeGPR: Deep Guided Posterior Regularization for Multi-Class Cell Detection and Counting,” Apr. 2023, Accessed: Jun. 08, 2023. [Online]. Available: https://arxiv.org/abs/2304.00741v1,

[86] S. Sadhukhan, G. K. Ghorai, S. Maiti, G. Sarkar, and A. K. Dhara, “Optic Disc Localization in Retinal Fundus Images using Faster R-CNN,” in *2018 Fifth International Conference on Emerging Applications of Information Technology (EAIT)*, Jan. 2018, pp. 1–4. [Online]. Available: https://ieeexplore.ieee.org/document/8470435/, doi:10.1109/EAIT.2018.8470435.

[87] S. Ajitha and M. V Judy, “Faster R-CNN classification for the recognition of glaucoma,” *J. Phys. Conf. Ser.*, vol. 1706, no. 1, p. 012170, Dec. 2020, [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1706/1/012170, doi:10.1088/1742-6596/1706/1/012170.

[88] G. Li, C. Li, C. Zeng, P. Gao, and G. Xie, “Region Focus Network for Joint Optic Disc and Cup Segmentation,” *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 01, pp. 751–758, Apr. 2020, [Online]. Available: https://aaai.org/ojs/index.php/AAAI/article/view/5418,

doi:10.1609/aaai.v34i01.5418.

[89] P. Kakade, A. Kale, I. Jawade, R. Jadhav, and N. Kulkarni, “Optic Disc Detection using Image Processing and Deep Learning,” vol. 3, no. 3, pp. 1–8, 2016, Accessed: Jun. 18, 2022. [Online]. Available: https://journal.uob.edu.bh/bitstream/handle/123456789/4435/1570702812.pdf?sequence=1,

[90] T. Nazir, A. Irtaza, and V. Starovoitov, “Optic Disc and Optic Cup Segmentation for Glaucoma Detection from Blur Retinal Images Using Improved Mask-RCNN,” *Int. J. Opt.*, vol. 2021, no. Dl, pp. 1–12, Jul. 2021, [Online]. Available: https://www.hindawi.com/journals/ijo/2021/6641980/, doi:10.1155/2021/6641980.

[91] Y. Guo, Y. Peng, and B. Zhang, “CAFR-CNN: coarse-to-fine adaptive faster R-CNN for cross-domain joint optic disc and cup segmentation,” *Appl. Intell.*, vol. 51, no. 8, pp. 5701–5725, Aug. 2021, [Online]. Available: https://link.springer.com/10.1007/s10489-020-02145-w, doi:10.1007/s10489-020-02145-w.

[92] H. Almubarak, Y. Bazi, and N. Alajlan, “Two-Stage Mask-RCNN Approach for Detecting and Segmenting the Optic Nerve Head, Optic Disc, and Optic Cup in Fundus Images,” *Appl. Sci.*, vol. 10, no. 11, p. 3833, May 2020, Accessed: May 12, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/10/11/3833, doi:10.3390/app10113833.

[93] Z. Wang, N. Dong, S. D. Rosario, M. Xu, P. Xie, and E. P. Xing, “ELLIPSE DETECTION OF OPTIC DISC-AND-CUP BOUNDARY IN FUNDUS IMAGES Zeya Wang , Nanqing Dong , Sean D Rosario , Min Xu , Pengtao Xie , Eric P . Xing,” *2019 IEEE 16th Int. Symp. Biomed. Imaging (ISBI 2019)*, no. Isbi, pp. 601–604, 2019.

[94] C. K. Lu, T. B. Tang, A. F. Murray, A. Laude, and B. Dhillon, “Automatic parapapillary atrophy shape detection and quantification in colour fundus

images," *2010 IEEE Biomed. Circuits Syst. Conf. BioCAS 2010*, pp. 86–89, 2010,doi:10.1109/BIOCAS.2010.5709577.

[95] C.-K. Lu, T. B. Tang, A. Laude, B. Dhillon, and A. F. Murray, "Parapapillary atrophy and optic disc region assessment (PANDORA): retinal imaging tool for assessment of the optic disc and parapapillary atrophy," *J. Biomed. Opt.*, vol. 17, no. 10, p. 1060101, Oct. 2012, Accessed: Jun. 08, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/23224009/, doi:10.1117/1.JBO.17.10.106010.

[96] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9351, pp. 234–241, 2015,doi:10.1007/978-3-319-24574-4_28.

[97] Y. Chai, H. Liu, and J. Xu, "A new convolutional neural network model for peripapillary atrophy area segmentation from retinal fundus images," *Appl. Soft Comput. J.*, vol. 86, Jan. 2020, Accessed: Sep. 15, 2023. [Online]. Available: https://dl.acm.org/doi/10.1016/j.asoc.2019.105890, doi:10.1016/J.ASOC.2019.105890.

[98] C. Wan *et al.*, "Optimized-Unet: Novel Algorithm for Parapapillary Atrophy Segmentation," *Front. Neurosci.*, vol. 15, Oct. 2021, Accessed: Jun. 08, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/34720868/, doi:10.3389/FNINS.2021.758887.

[99] Y. Chai, H. Liu, and J. Xu, "Glaucoma diagnosis based on both hidden features and domain knowledge through deep learning models," *Knowledge-Based Syst.*, vol. 161, pp. 147–156, Dec. 2018,doi:10.1016/J.KNOSYS.2018.07.043.

[100] M. Mateen, T. S. Malik, S. Hayat, M. Hameed, S. Sun, and J. Wen, "Deep Learning Approach for Automatic Microaneurysms Detection," *Sensors 2022, Vol. 22, Page 542*, vol. 22, no. 2, p. 542, Jan. 2022, Accessed: Jun. 08, 2023.

[Online]. Available: https://www.mdpi.com/1424-8220/22/2/542/htm, doi:10.3390/S22020542.

[101] G. B. Kande, T. S. Savithri, and P. V. Subbaiah, “Automatic Detection of Microaneurysms and Hemorrhages in Digital Fundus Images,” *J. Digit. Imaging*, vol. 23, no. 4, p. 430, Aug. 2010, Accessed: Jun. 08, 2023. [Online]. Available: /pmc/articles/PMC3046669/, doi:10.1007/S10278-009-9246-0.

[102] B. Borsos, L. Nagy, D. Iclănzan, and L. Szilágyi, “Automatic detection of hard and soft exudates from retinal fundus images,” *Acta Univ. Sapientiae, Inform.*, vol. 11, no. 1, pp. 65–79, Aug. 2019,doi:10.2478/AUSI-2019-0005.

[103] C. Santos, M. Aguiar, D. Welfer, and B. Belloni, “A New Approach for Detecting Fundus Lesions Using Image Processing and Deep Neural Network Architecture Based on YOLO Model,” *Sensors*, vol. 22, no. 17, p. 6441, Sep. 2022, Accessed: Jun. 08, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/22/17/6441/htm, doi:10.3390/S22176441/S1.

[104] L. Dai *et al.*, “A deep learning system for detecting diabetic retinopathy across the disease spectrum,” *Nat. Commun. 2021 121*, vol. 12, no. 1, pp. 1–11, May 2021, Accessed: Jun. 08, 2023. [Online]. Available: https://www.nature.com/articles/s41467-021-23458-5, doi:10.1038/s41467-021-23458-5.

[105] T. Nazir *et al.*, “Detection of Diabetic Eye Disease from Retinal Images Using a Deep Learning Based CenterNet Model,” *Sensors 2021, Vol. 21, Page 5283*, vol. 21, no. 16, p. 5283, Aug. 2021, Accessed: Jun. 08, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/21/16/5283/htm, doi:10.3390/S21165283.

[106] W. L. Alyoubi, M. F. Abulkhair, and W. M. Shalash, “Diabetic Retinopathy Fundus Image Classification and Lesions Localization System Using Deep Learning,” *Sensors 2021, Vol. 21, Page 3704*, vol. 21, no. 11, p. 3704, May

2021, Accessed: Jun. 08, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/21/11/3704/htm, doi:10.3390/S21113704.

[107] N. Mahabadi and Y. Al Khalili, “Neuroanatomy, Retina,” *StatPearls*, Aug. 2022, Accessed: Jun. 08, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK545310/,

[108] J. Son, J. Y. Shin, H. D. Kim, K. H. Jung, K. H. Park, and S. J. Park, “Development and Validation of Deep Learning Models for Screening Multiple Abnormal Findings in Retinal Fundus Images,” *Ophthalmology*, vol. 127, no. 1, pp. 85–94, Jan. 2020, Accessed: Jan. 17, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31281057/, doi:10.1016/J.OPHTHA.2019.05.029.

[109] J. Wang, L. Yang, Z. Huo, W. He, and J. Luo, “Multi-Label Classification of Fundus Images with EfficientNet,” *IEEE Access*, vol. 8, pp. 212499–212508, 2020,doi:10.1109/ACCESS.2020.3040275.

[110] S. Karthikeyan, P. Sanjay Kumar, R. J. Madhusudan, S. K. Sundaramoorthy, and P. K. Krishnan Namboori, “Detection of multi-class retinal diseases using artificial intelligence: An expeditious learning using deep CNn with minimal data,” *Biomed. Pharmacol. J.*, vol. 12, no. 3, pp. 1577–1586, 2019,doi:10.13005/BPJ/1788.

[111] D. S. W. Ting *et al.*, “Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes,” *JAMA*, vol. 318, no. 22, pp. 2211–2223, Dec. 2017, Accessed: Dec. 19, 2022. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29234807/, doi:10.1001/JAMA.2017.18152.

[112] L. P. Cen *et al.*, “Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks,” *Nat. Commun. 2021 121*, vol. 12, no. 1, pp. 1–13, Aug. 2021, Accessed: Jun. 08, 2023. [Online]. Available: https://www.nature.com/articles/s41467-021-25138-w,

doi:10.1038/s41467-021-25138-w.

[113] T. H. Rim *et al.*, “Prediction of systemic biomarkers from retinal photographs: development and validation of deep-learning algorithms,” *Lancet Digit. Heal.*, vol. 2, no. 10, pp. e526–e536, Oct. 2020, Accessed: Jan. 17, 2023. [Online]. Available: http://www.thelancet.com/article/S2589750020302168/fulltext, doi:10.1016/S2589-7500(20)30216-8.

[114] T.-Y. Lin *et al.*, “Microsoft COCO: Common Objects in Context,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 3686–3693, May 2014, Accessed: Apr. 01, 2022. [Online]. Available: http://arxiv.org/abs/1405.0312,

[115] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The Pascal Visual Object Classes (VOC) Challenge,” *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010, [Online]. Available: http://link.springer.com/10.1007/s11263-009-0275-4, doi:10.1007/s11263-009-0275-4.

[116] Z. Cai and N. Vasconcelos, “Cascade R-CNN: High Quality Object Detection and Instance Segmentation,” Jun. 2019, Accessed: Mar. 21, 2022. [Online]. Available: http://arxiv.org/abs/1906.09756,

[117] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, “Mask Scoring R-CNN,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, vol. 2019-June, pp. 6402–6411. [Online]. Available: https://ieeexplore.ieee.org/document/8953609/, doi:10.1109/CVPR.2019.00657.

[118] A. Kirillov, Y. Wu, K. He, and R. Girshick, “PointRend: Image Segmentation as Rendering,” Dec. 2019, [Online]. Available: http://arxiv.org/abs/1912.08193,

[119] J. Wang, K. Chen, R. Xu, Z. Liu, C. C. Loy, and D. Lin, “CARAFE: Content-Aware ReAssembly of FEatures,” May 2019, Accessed: Jun. 18, 2022.

[Online]. Available: http://arxiv.org/abs/1905.02188,

[120] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, “GCNet: Non-local Networks Meet Squeeze-Excitation Networks and Beyond,” Apr. 2019, Accessed: Jun. 18, 2022. [Online]. Available: http://arxiv.org/abs/1904.11492,

[121] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li, “SOLO: Segmenting Objects by Locations,” Dec. 2019, Accessed: Mar. 24, 2022. [Online]. Available: http://arxiv.org/abs/1912.04488, doi:https://doi.org/10.48550/arXiv.1912.04488.

[122] A. Dutta and A. Zisserman, “The VIA annotation software for images, audio and video,” *MM 2019 - Proc. 27th ACM Int. Conf. Multimed.*, pp. 2276–2279, Oct. 2019,doi:10.1145/3343031.3350535.

[123] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-December, pp. 770–778, Dec. 2015, Accessed: Oct. 21, 2022. [Online]. Available: https://arxiv.org/abs/1512.03385v1, doi:10.48550/arxiv.1512.03385.

[124] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature Pyramid Networks for Object Detection,” Dec. 2016, Accessed: May 29, 2022. [Online]. Available: http://arxiv.org/abs/1612.03144,

[125] G. Peyré and M. Cuturi, “Computational Optimal Transport,” *Found. Trends Mach. Learn.*, vol. 11, no. 5–6, pp. 1–257, Mar. 2018, Accessed: Jun. 19, 2023. [Online]. Available: https://arxiv.org/abs/1803.00567v4, doi:10.1561/2200000073.

[126] A. Rame, E. Garreau, H. Ben-Younes, and C. Ollion, “OMNIA Faster R-CNN: Detection in the wild through dataset merging and soft distillation,” Dec. 2018, Accessed: Jun. 14, 2023. [Online]. Available: https://arxiv.org/abs/1812.02611v2,

[127] “OpenCV - Open Computer Vision Library.” https://opencv.org/ (accessed Jun. 15, 2023).

[128] J. Günther, P. M. Pilarski, G. Helfrich, H. Shen, and K. Diepold, “First Steps Towards an Intelligent Laser Welding Architecture Using Deep Neural Networks and Reinforcement Learning,” *Procedia Technol.*, vol. 15, pp. 474–483, 2014, [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2212017314001224, doi:10.1016/j.protcy.2014.09.007.

[129] K. Oksuz, B. C. Cam, S. Kalkan, and E. Akbas, “Imbalance Problems in Object Detection: A Review.,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3388–3415, Mar. 2020, Accessed: Jun. 16, 2023. [Online]. Available: https://open.metu.edu.tr/handle/11511/70222, doi:10.1109/TPAMI.2020.2981890.

[130] A. Shrivastava, A. Gupta, and R. Girshick, “Training Region-based Object Detectors with Online Hard Example Mining”.

[131] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, “Path Aggregation Network for Instance Segmentation,” Mar. 2018, Accessed: Jun. 18, 2022. [Online]. Available: http://arxiv.org/abs/1803.01534,

[132] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, “UnitBox: An advanced object detection network,” *MM 2016 - Proc. 2016 ACM Multimed. Conf.*, pp. 516–520, Oct. 2016, Accessed: Jun. 17, 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/2964284.2967274, doi:10.1145/2964284.2967274.

[133] L. Tychsen-Smith and L. Petersson, “Improving Object Localization with Fitness NMS and Bounded IoU Loss,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 6877–6885, Nov. 2017, Accessed: Jun. 17, 2023. [Online]. Available: https://arxiv.org/abs/1711.00164v3, doi:10.1109/CVPR.2018.00719.

[134] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 658–666, Jun. 2019,doi:10.1109/CVPR.2019.00075.

[135] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 07, pp. 12993–13000, Apr. 2020, Accessed: Jun. 17, 2023. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/6999, doi:10.1609/AAAI.V34I07.6999.

[136] C. Xu, J. Wang, W. Yang, H. Yu, L. Yu, and G.-S. Xia, "Detecting tiny objects in aerial images: A normalized Wasserstein distance and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 190, pp. 79–93, Jun. 2022, Accessed: Jun. 17, 2023. [Online]. Available: http://arxiv.org/abs/2206.13996, doi:10.1016/j.isprsjprs.2022.06.002.

[137] J. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 6469–6477, May 2017, Accessed: Jun. 19, 2023. [Online]. Available: https://arxiv.org/abs/1705.02950v2, doi:10.1109/CVPR.2017.685.

[138] Z. Zheng *et al.*, "Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance Segmentation," *IEEE Trans. Cybern.*, vol. 52, no. 8, pp. 8574–8586, May 2020, Accessed: Jun. 19, 2023. [Online]. Available: https://arxiv.org/abs/2005.03572v4, doi:10.1109/TCYB.2021.3095305.

[139] W. Y. Lee, S. M. Park, and K. B. Sim, "Optimal hyperparameter tuning of convolutional neural networks based on the parameter-setting-free harmony search algorithm," *Optik (Stuttg).*, vol. 172, pp. 359–367, Nov. 2018,doi:10.1016/J.IJLEO.2018.07.044.

[140] J. Bergstra, J. B. Ca, and Y. B. Ca, “Random search for hyper-parameter optimization,” *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, Feb. 2012, Accessed: Jul. 04, 2023. [Online]. Available: https://dl.acm.org/doi/10.5555/2188385.2188395, doi:10.5555/2188385.2188395.

[141] Q. Huang, J. Mao, and Y. Liu, “An improved grid search algorithm of SVR parameters optimization,” *Int. Conf. Commun. Technol. Proceedings, ICCT*, pp. 1022–1026, 2012,doi:10.1109/ICCT.2012.6511415.

[142] D. Maclaurin, D. Duvenaud, R. P. Adams, D. Maclaurin, D. Duvenaud, and R. P. Adams, “Gradient-based Hyperparameter Optimization through Reversible Learning,” *arXiv*, p. arXiv:1502.03492, Feb. 2015, Accessed: Jul. 04, 2023. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2015arXiv150203492M/abstract, doi:10.48550/ARXIV.1502.03492.

[143] I. Loshchilov and F. Hutter, “DECOUPLED WEIGHT DECAY REGULARIZATION.”doi:https://doi.org/10.48550/arXiv.1711.05101.

[144] I. Loshchilov and F. Hutter, “SGDR: Stochastic Gradient Descent with Warm Restarts,” Aug. 2016, Accessed: Apr. 19, 2022. [Online]. Available: http://arxiv.org/abs/1608.03983, doi:https://doi.org/10.48550/arXiv.1608.03983.

[145] S. Norouzi and M. Ebrahimi, “A Survey on Proposed Methods to Address Adam Optimizer Deficiencies.” Accessed: Jun. 18, 2022. [Online]. Available: http://www.cs.toronto.edu/~sajadn/sajad_norouzi/ECE1505.pdf,

[146] J. Patterson and A. Gibson, “Deep Learning A Practioner’s Approach,” *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1–994, 2017.

[147] T. Ridnik *et al.*, “Asymmetric Loss For Multi-Label Classification,” *2021 IEEE/CVF Int. Conf. Comput. Vis.*, pp. 82–91, Oct. 2021,doi:10.1109/ICCV48922.2021.00015.

[148] Z. Zhang and M. R. Sabuncu, “Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels”, Accessed: Jul. 04, 2023. [Online]. Available: https://dl.acm.org/doi/10.5555/3327546.3327555, doi:10.5555/3327546.3327555.

[149] D. Müller, I. Soto-Rey, and F. Kramer, “Towards a Guideline for Evaluation Metrics in Medical Image Segmentation,” pp. 1–7, Feb. 2022, [Online]. Available: http://arxiv.org/abs/2202.05273, doi:https://doi.org/10.48550/arXiv.2202.05273.

[150] D. Bolya, S. Foley, J. Hays, and J. Hoffman, “TIDE: A General Toolbox for Identifying Object Detection Errors,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12348 LNCS, pp. 558–573, 2020, Accessed: Jul. 07, 2023. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-58580-8_33, doi:10.1007/978-3-030-58580-8_33/COVER.

[151] “COCO - Common Objects in Context.” https://cocodataset.org/#detection-eval (accessed Apr. 22, 2022).

[152] E. J. Carmona, M. Rincón, J. García-Feijoó, and J. M. Martínez-de-la-Casa, “Identification of the optic nerve head with genetic algorithms,” *Artif. Intell. Med.*, vol. 43, no. 3, pp. 243–259, Jul. 2008, [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0933365708000547, doi:10.1016/j.artmed.2008.04.005.

[153] A. Arpteg, B. Brinne, L. Crnkovic-Friis, and J. Bosch, “Software engineering challenges of deep learning,” *Proc. - 44th Euromicro Conf. Softw. Eng. Adv. Appl. SEAA 2018*, no. August, pp. 50–59, 2018,doi:10.1109/SEAA.2018.00018.

[154] Z. Chen, Y. Cao, Y. Liu, H. Wang, T. Xie, and X. Liu, “A Comprehensive Study on Challenges in Deploying Deep Learning Based Software,” *ESEC/FSE 2020 - Proc. 28th ACM Jt. Meet. Eur. Softw. Eng. Conf. Symp. Found. Softw.*

Eng., pp. 750–762, May 2020, Accessed: Nov. 23, 2023. [Online]. Available: https://arxiv.org/abs/2005.00760v4, doi:10.1145/3368089.3409759.

[155] R. V. Shahir Daya Nguyen Van Duy, Kameswara Eati, Carlos M Ferreira, Dejan Glozic, Vasfi Gucer, Manav Gupta, Sunil Joshi, Valerie Lampkin, Marcelo martins, Shishir Narain, "Microservices from Theory to Practice Creating Applications in IBM Bluemix Using the Microservices Approach," *Ibm*, p. 170, 2015.

[156] C. Boettiger, "An introduction to Docker for reproducible research," *ACM SIGOPS Oper. Syst. Rev.*, vol. 49, no. 1, pp. 71–79, Jan. 2015, Accessed: Nov. 24, 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/2723872.2723882, doi:10.1145/2723872.2723882.

[157] A. Javeed, "Performance Optimization Techniques for ReactJS," *Proc. 2019 3rd IEEE Int. Conf. Electr. Comput. Commun. Technol. ICECCT 2019*, Feb. 2019,doi:10.1109/ICECCT.2019.8869134.

[158] S. Shahinfar, P. Meek, and G. Falzon, "'How many images do I need?' Understanding how sample size per class affects deep learning model performance metrics for balanced designs in autonomous wildlife monitoring," *Ecol. Inform.*, vol. 57, p. 101085, May 2020, Accessed: Jun. 01, 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1574954120300352, doi:10.1016/j.ecoinf.2020.101085.

[159] D. Hoiem, Y. Chodpathumwan, and Q. Dai, "Diagnosing Error in Object Detectors," in *Computer Vision -- ECCV 2012*, 2012, pp. 340–353.

[160] L. Gao, Y. He, X. Sun, X. Jia, and B. Zhang, "Incorporating Negative Sample Training for Ship Detection Based on Deep Learning," *Sensors*, vol. 19, no. 3, p. 684, Feb. 2019, Accessed: Jun. 16, 2022. [Online]. Available: http://www.mdpi.com/1424-8220/19/3/684, doi:10.3390/s19030684.

[161] J.-B. Grill *et al.*, “Bootstrap your own latent: A new approach to self-supervised Learning,” *CoRR*, vol. abs/2006.0, Jun. 2020, [Online]. Available: http://arxiv.org/abs/2006.07733,

[162] “Causal Models (Stanford Encyclopedia of Philosophy).” https://plato.stanford.edu/entries/causal-models/ (accessed Oct. 26, 2020).

[163] J. Pearl, *Causality: Models, reasoning, and inference, second edition*. 2011.doi:10.1017/CBO9780511803161.