

Título del trabajo:

Desarrollo de un modelo de inteligencia artificial nacional para el tamizado de casos de retinopatía diabética referible a partir de fotografías de fondo de ojo

Nombre del premio al que aspira:

Premio anual “ACADEMIA NACIONAL DE MEDICINA”.

Nombre y cargo de los autores

- Dr. José Ignacio Orlando (Investigador Asistente CONICET / Grupo Yatiris, Instituto PLADEMA, UNICEN).
- Ing. Tomás Castilla (Grupo Yatiris, Instituto PLADEMA, UNICEN).
- Dr. Alejandro Koch (Servicio de Cirugía Cardiovascular, Hospital de Alta Complejidad En Red “El Cruce” Dr. Néstor Carlos Kirchner).
- Dra. Marcela S. Martínez (Centro de Oftalmología Martínez).
- Dra. Mercedes Leguía (Jefa del Servicio de Oftalmología del Hospital de Alta Complejidad En Red “El Cruce” Dr. Néstor Carlos Kirchner).
- Dr. Ignacio Larrabide (Investigador Independiente CONICET / Coordinador del Grupo Yatiris, Instituto PLADEMA, UNICEN).

Instituciones participantes

- *Lugar donde fue realizado el trabajo:* Grupo Yatiris, Instituto PLADEMA, Fac. de Cs. Exactas, Universidad Nacional del Centro de la Provincia de Buenos Aires (UNICEN) (Tandil, Argentina).
- *Instituciones colaboradoras:* Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Servicio de Oftalmología, Hospital de Alta Complejidad En Red “El Cruce” Dr. Néstor Carlos Kirchner (Florencio Varela, Argentina), Centro Oftalmológico Martínez (Pehuajó, Argentina).

Fecha de finalización del trabajo

31 de marzo de 2022 (este trabajo no ha sido publicado aún)

Resumen

Los modelos de inteligencia artificial (IA) han permitido mejorar la eficiencia de los sistemas de telemedicina oftalmológica para tamizado temprano de la retinopatía diabética (RD) mediante la detección automática de los casos de riesgo, lo que reduce la carga de trabajo de los oftalmólogos que realizan los informes y les permite dedicar mayor tiempo al tratamiento de los casos urgentes. Sin embargo, Argentina no cuenta con herramientas propias para este fin, viéndose obligada a importar costosas licencias comerciales y tener que adaptarse al diseño rígido de las mismas.

El objetivo de este trabajo es introducir un modelo de IA nacional para la detección automática de casos de RD referible a partir de fotografías de fondo de ojo (FFO), que pueda ejecutarse con éxito sobre imágenes obtenidas con cualquier dispositivo de captura, para facilitar su aplicación en centros médicos de Argentina sin necesidad de cambiar su aparatología actual.

El modelo propuesto consiste en una red neuronal convolucional con arquitectura residual de 18 capas, entrenada sobre 39.592 retinografías preprocesadas automáticamente para incrementar su contraste y resaltar potenciales lesiones patológicas. Las mismas fueron extraídas de 5 particiones de entrenamiento provistas en las bases de datos públicas DDR, DeepDRID, APTOS2019 y EyePACS, y fueron clasificadas en las categorías “sin RD referible” y “con RD referible” a partir de anotaciones propias del grado de RD según la ETDRS. El modelo fue entrenado sobre estos datos para reproducir dichas etiquetas.

El algoritmo resultante se evaluó mediante múltiples protocolos sobre un total de 61.526 estudios provenientes de 11 conjuntos de imágenes diferentes, 9 de acceso público (DDR, DeepDRID, APTOS2019, EyePACS, IDRiD, FCM-UNA, 1000Fundus, DR2, MESSIDOR2 y DIARETDB1) y 2 obtenidos retrospectivamente de las bases de datos clínicas del Centro Oftalmológico Martínez (Pehuajó, Argentina) y del Hospital El Cruce (HEC) y etiquetados manualmente por dos oftalmólogos especialistas. Ninguna de estas imágenes fue empleada para el entrenamiento del algoritmo.

Para detectar RD referible a partir de una única imagen, el modelo reportó valores de área bajo curva ROC (AUC), sensibilidad (Se) y especificidad (Esp) en los conjuntos públicos más popularmente utilizados en línea o superiores a los de otros trabajos de la literatura, a saber: en los datos de test de EyePACS (N = 53.576 imágenes): AUC = 0.951 (IC 95%, 0.949-0.954), Se = 73.2% (IC 95%, 72.3%-74%), Esp = 97.9% (IC 95%, 97.8%-98.0%); en los de test de DDR (N = 3.759): AUC = 0.965 (IC 95%, 0.960-0.970), Se = 74.9% (IC 95%, 72.8%-76.9%), Esp = 97.8% (IC 95%, 97.1%-98.4%); en los de MESSIDOR (N = 1.744): AUC = 0.973 (IC 95%, 0.967 - 0.979), Se = 89.5% (IC 95%, 86.3%-92.0%), Esp = 94.1% (IC 95%, 92.7%-95.3%); en los de DR2 (N = 435): AUC = 0.974 (IC 95%, 0.962-0.985), Se = 84.7% (IC 95%, 76.3%-90.5%), Esp = 96.1% (IC 95%, 93.5%-97.7%); y en los de test de IDRiD (N = 103): AUC = 0.949 (IC 95%, 0.914-0.980), Se = 82.8% (IC 95%, 71.8%-90.1%), Esp = 89.7% (IC 95%, 76.4%-95.9%). Además, en datos nacionales del Centro Oftalmológico Martínez (N = 484), el modelo reportó AUC = 0.955 (IC 95%, 0.927-0.980), Se = 80.0% (IC 95%, 62.7%-90.5%), Esp = 93.4% (IC 95%, 90.7%-95.3%). En imágenes del HEC (N = 35) se obtuvieron valores de AUC = 0.961 (IC 95%, 0.900 - 1.000), Se = 100% (IC 95%, 61.0%-100%) y Esp = 86.2% (IC 95%, 69.4%-94.5%). Se realizaron también evaluaciones combinando múltiples estudios por ojo del paciente, desagrupando por calidad de imagen, grado de RD y sobre imágenes con patologías u observaciones alternativas, diferentes de la RD. Según el relevamiento de la literatura realizado, es el primer trabajo en evaluar extensivamente en tan múltiples contextos un modelo de inteligencia artificial para reconocimiento de RD. Los resultados obtenidos sugieren que el modelo es lo suficientemente robusto para aplicarse al filtrado de casos de RD referible en el contexto de redes de teleoftalmología. En el mediano plazo se espera integrar este algoritmo en una herramienta nacional que pueda aplicarse para llevar el diagnóstico temprano de la enfermedad a pacientes diabéticos que aún no cumplen con el control oftalmológico anual recomendado.

1. Introducción

La retinopatía diabética (RD) es la primera causa de ceguera prevenible e irreversible en adultos en edad laboral [1]. Su detección a tiempo es clave para que los tratamientos sean efectivos, por lo que se recomienda a la persona diabética realizarse un control de fondo de ojos (retina-vítreo) al menos una vez al año. Sin embargo, la disponibilidad limitada de oftalmólogos y su concentración geográfica hacen que sólo un 30% de los diabéticos cumpla con el control [2], en muchos casos recurriendo a la consulta ya en estado de ceguera legal y sin posibilidad alguna de ser tratados eficazmente. La telemedicina oftalmológica ofrece una alternativa para movernos de esta dificultad. A través de la instalación de nodos de captura en lugares estratégicos, personal no-oftalmológico con un entrenamiento básico puede adquirir fotografías de fondo de ojo (FFO) de personas de riesgo, y transmitirlas a un centro de informes. Allí, un grupo de oftalmólogos establece el diagnóstico correspondiente y, en caso de observar signos de la enfermedad, recomienda el tratamiento más conveniente [3]. Esto es posible gracias a la facilidad de captura de la FFO, que se obtiene de manera no-invasiva y mediante dispositivos con un costo asociado relativamente bajo. Sin embargo, estas redes tienen una capacidad de escalabilidad limitada, ya que el incremento en la afluencia de personas diabéticas a los nodos o la incorporación de muchos nodos nuevos puede saturar al centro de informes, obligando a los profesionales que realizan el diagnóstico a dedicar más tiempo al estudio de las imágenes recibidas que al tratamiento efectivo de los pacientes.

Los algoritmos de visión computacional basados en inteligencia artificial (IA) son hoy día capaces de clasificar imágenes naturales de manera automática con una efectividad comparable con la humana [4]. Su aplicación en medicina se ha popularizado notoriamente en los últimos años, empleándose para tareas tales como la detección automática de enfermedades [5,6], la asistencia a la planificación de tratamientos [7] y el seguimiento posterior de los mismos [8], o hasta para el

descubrimiento de biomarcadores para el diseño de nuevas estrategias terapéuticas [9]. En el campo específico de la telemedicina oftalmológica, algunos desarrollos comerciales recientes en Europa, Estados Unidos y Asia usan estas tecnologías para reducir la carga de trabajo de los centros de informes, detectando y luego filtrando automáticamente los casos que requieren ser analizados por un profesional mediante IA [10]. Sin embargo, estos modelos requieren usar cámaras específicas o están entrenados para etnias y poblaciones determinadas, por lo que fallan al aplicarse sobre estudios de otros orígenes [11]. Así mismo, no existen en Argentina desarrollos nacionales en esta dirección, lo que hoy por hoy detiene la posibilidad de planificar estrategias de tamizado poblacional masivo utilizando los recursos disponibles, obligando al país a adquirir estas tecnologías a un elevado costo y a tener que adaptarse a las condiciones impuestas por estas herramientas rígidas.

En este trabajo se presenta la implementación nacional y evaluación de un modelo de aprendizaje profundo basado en redes neuronales convolucionales para el reconocimiento automático de la RD a partir de FFO, desarrollado íntegramente en Argentina con el propósito de facilitar la costo-efectividad del tamizado de la enfermedad en grandes poblaciones de riesgo. El algoritmo toma como entrada una FFO capturada con cualquier cámara midriática o no-midriática y predice un valor de probabilidad que indica qué tan factible es que el paciente presente un nivel de RD que requiera de una consulta médica (RD referible). Para favorecer su aplicabilidad en nuestro país, el método fue entrenado con imágenes capturadas con múltiples dispositivos y provenientes de diversas poblaciones mundiales, recolectadas masivamente a partir de un relevamiento exhaustivo de bancos de fotografías públicas y de acceso abierto. Mediante la incorporación de técnicas de preprocesamiento y de un entrenamiento cuidadoso, el modelo final es capaz de producir respuestas precisas, estables y robustas ante diferentes escenarios de evaluación. En particular, se realizó un estudio de la performance del método utilizando más de 61.000 FFO, analizando su sensibilidad y especificidad en

múltiples cohortes y comprobando su efectividad para reconocer diferentes grados de RD y para trabajar con imágenes de calidad variable o pertenecientes a pacientes con otras patologías. Por otro lado, se recolectaron imágenes de manera retrospectiva en instituciones de salud de Argentina para estudiar los resultados que el algoritmo puede obtener sobre estudios nacionales, comprobando su capacidad para identificar correctamente los casos con RD referible.

2. Explicación de los fundamentos

Según estadísticas del Banco Mundial, la prevalencia de la diabetes en personas de entre 20 y 79 años en Argentina fue del 5.7% en el año 2010, y del 5.9% en el año 2019.¹ De acuerdo al INDEC, se proyecta que en 2022 la población argentina dentro de esta franja etaria ascenderá a 30.306.130 personas.² Asumiendo un comportamiento lineal creciente de la diabetes, en acuerdo con las proyecciones internacionales de la Organización Mundial de la Salud,³ en 2022 se espera que esta enfermedad afecte al 5.97% de esta población. Puede estimarse entonces que la cantidad de diabéticos en edad laboral y/o en la primera etapa de la tercera edad ascenderá en 2022 a 1.780.040 personas. Toda esta población requiere indefectiblemente realizarse al menos un estudio oftalmológico anual para monitorear el estado de sus retinas y poder detectar a tiempo si presentan signos de RD que requieran tratamiento para reducir el riesgo de ceguera. Sin embargo, en Argentina se estima que existen solamente unos 4.500 oftalmólogos [12], un número que resulta insuficiente para cubrir el tamaño de esta demanda. Tal es así que sería necesario que cada oftalmólogo del país atienda diariamente a más de una persona diabética para abastecer esta necesidad, algo infactible si se tiene en cuenta que existen otras patologías oculares que también requieren atención. Por otro lado,

¹ <https://data.worldbank.org/indicator/SH.STA.DIAB.ZS?locations=ZJ-AR>

²

https://www.indec.gob.ar/ftp/cuadros/publicaciones/proyeccionesyestimaciones_nac_2010_2040.pdf

³ <https://www.who.int/news-room/fact-sheets/detail/diabetes>

estadísticas recientes han demostrado que el nivel de adhesión del paciente diabético al control anual oftalmológico es cercano al 30% [2]. En consecuencia, estas personas concurren a la consulta cuando comienzan a experimentar síntomas severos y ya es demasiado tarde para recuperar la visión perdida. Según un estudio retrospectivo realizado por el Servicio de Oftalmología del Hospital El Cruce (Florencio Varela, Argentina), un 53% de los casos de RD del hospital registrados entre 2017 y 2018 presentaba en su primera consulta signos de RD severa o proliferativa, y un 20% correspondía ya a casos de ceguera legal [13].

Algunas de las causas observadas de la falta de adhesión al control oftalmológico anual incluyen una educación pobre acerca de la gravedad de la patología, el alto costo de acceso a la consulta y la concentración mayoritaria de los profesionales en grandes centros urbanos. En Argentina, por ejemplo, se ha observado que un gran número de los oftalmólogos disponibles reside o atiende en las grandes urbes [14], aunque esta distribución no es directamente proporcional a la densidad poblacional de estos núcleos urbanos [13]. Estas últimas dos dificultades pueden aliviarse mediante telemedicina, instalando nodos de captura de FFO en las zonas sin cobertura oftalmológica. Así, técnicos con una capacitación básica en captura de imágenes puede adquirir a bajo costo estudios de pacientes diabéticos que habitan en regiones con baja o nula disponibilidad de profesionales, y transmitirlos hacia centros de informe para que sean analizados por los oftalmólogos.

Este tipo de iniciativas han demostrado reducir riesgos y costos para el tamizado de la enfermedad [3]. Sin embargo, pierden efectividad cuando el número de nodos de captura crece en demasia o cuando la cantidad de pacientes a analizar es demasiado grande, ya que esto provoca un incremento significativo en la carga de trabajo del centro de informes, requiriendo más cantidad de profesionales para el diagnóstico y afectando su disponibilidad para otras tareas de relevancia tales como tratar pacientes cuya patología retinal requiere de una intervención rápida.

El tamizado automático mediante inteligencia artificial de aquellos casos que requieren ser verificados por profesionales permitiría evitar la transmisión al nodo de informes de aquellos estudios que corresponden a personas diabéticas sin signos de progreso de RD, reduciendo la cantidad de imágenes a analizar y asegurando un aprovechamiento más efectivo del recurso humano. Numerosos países se han inclinado por este tipo de soluciones en los últimos años [14, 15]. Sin embargo, nuestro país no cuenta con tecnologías propias en este área, por lo que para acceder a ellas debe adquirir softwares comerciales que a menudo requieren un dispositivo de captura específico o están diseñados para un sistema sanitario en particular [10].

3. Objetivos del trabajo

En este trabajo se presenta el desarrollo de un modelo de IA basado en aprendizaje profundo que utiliza una red neuronal convolucional para el reconocimiento automático de casos de RD que requieran de intervención oftalmológica, a partir de FFO. El objetivo del mismo es identificar automáticamente aquellas FFO que se corresponden con casos de RD no proliferativa (RDNP) moderada o severa o de RD proliferativa (RDP), respecto de casos sin signos de RD o con signos de RDNP leve. Para ello, el algoritmo asigna una probabilidad de referibilidad que, cuanto más cercana a 1 es, más refleja el grado de confianza de que se trate de un caso de riesgo. Además, el mismo genera un mapa de activaciones que permite estudiar las regiones de la imagen que fueron tenidas en cuenta para determinar la probabilidad correspondiente. En consecuencia, estos dos insumos permiten por un lado automatizar el tamizado de los casos de riesgo (reduciendo la carga de trabajo de los nodos de informe en redes de teleoftalmología) y al mismo tiempo retroalimentar el proceso de diagnóstico clínico tradicional, brindando información extra que facilita a los oftalmólogos el informe del estudio.

Para validar la efectividad del algoritmo propuesto, se introduce además un protocolo de evaluación detallado que permite estudiar las respuestas en numerosos

escenarios de interés, incluyendo evaluaciones de las predicciones realizadas sobre imágenes individuales, combinando múltiples estudios de un mismo paciente, utilizando diversas bases de datos públicas conocidas, comparando resultados con otros trabajos de la literatura, evaluando sobre imágenes recolectadas de manera retrospectiva en instituciones de salud de nuestro país, y estudiando la variabilidad de los resultados en estudios de distinta calidad, con grados diversos de RD o con otras patologías de interés.

4. Material y métodos

4.1. Bases de datos utilizadas

Para el entrenamiento, ajuste y validación del algoritmo propuesto se realizó un relevamiento exhaustivo de las bases de datos de FFO disponibles en acceso público. Se tuvieron en cuenta únicamente aquellas que contarán con etiquetas indicando si cada imagen presentaba o no RD referible, o que proveyeran o bien el grado de la RD según alguna escala clínica bien documentada o segmentaciones manuales de lesiones asociadas a la RD. Para generar etiquetas de referibilidad a partir de grados de RD de la ETDRS (0 - sin RD, 1 - RD no proliferativa leve, 2 - RD no proliferativa moderada, 3 - RD no proliferativa severa, 4 - RD proliferativa), se asociaron los grados 2, 3 y 4 a la clase “con RD referible”, y los grados 0 y 1 a la clase “sin RD referible”. En casos en los que se utilizara otra escala o segmentaciones, el criterio de etiquetado se estableció individualmente (ver siguientes subsecciones). Como resultado del relevamiento realizado, se identificaron 10 conjuntos de estudios por un total de 117.168 imágenes (ver Tabla 1). La Figura 1 presenta un resumen de algunas muestras de cada conjunto de datos público utilizado, representando por separado los casos sin y con RD referible. Nótese la variabilidad en grados de iluminación, artefactos, resoluciones y campo visual (o *field-of-view*, FOV) en cada uno de ellos.

Las imágenes recolectadas fueron separadas en conjuntos disjuntos de entrenamiento y validación (utilizados el uno para entrenar el modelo y el otro para calibrar sus hiperparámetros, respectivamente), y se construyeron además múltiples conjuntos también disjuntos, llamados de test, diseñados para evaluar el algoritmo en diferentes escenarios de interés. En líneas generales, se construyó un conjunto de entrenamiento con 39.592 estudios adquiridos a partir de particiones que en las bases de datos DDR, DeepDRID, APTOS2019 y EyePACS ya están predeterminados como datos de entrenamiento. El conjunto de validación consistió en 9.569 imágenes recolectadas en parte del conjunto de validación predeterminado de DDR y extraídas aleatoriamente de los predeterminados como entrenamiento en EyePACS e IDRiD. Finalmente, se establecieron 12 conjuntos diferentes para el testeo final de la red propuesta en diversas condiciones, que en total suman 61.526 imágenes. Téngase en cuenta que no hay imágenes repetidas en ninguno de los conjuntos. Las características individuales de cada base de datos y la estrategia aplicada para su etiquetado y su agrupamiento posterior se describen en las siguientes subsecciones.

4.1.1. DDR

DDR [17] cuenta con imágenes de 9.598 pacientes provenientes de 23 provincias distintas de China, con una edad promedio de 54 años (rango 1-100), todas centradas en la mácula y capturadas con 42 tipos diferentes de cámaras. Cada imagen se libera acompañada de una etiqueta de grado de RD según la escala ETDRS, o en su defecto indicando si la imagen no pudo ser evaluada por problemas en su calidad. Las imágenes sin diagnóstico asociado producto de la mala calidad de la imagen no fueron incluidas en este trabajo. Como el conjunto ya ofrece particiones de entrenamiento, validación y test predeterminadas, se respetaron las mismas para favorecer la comparación posterior de resultados con otros trabajos de la literatura.

4.1.2. DeepDRID

DeepDRID⁴ ofrece 2.000 imágenes de 500 pacientes capturadas con una cámara Topcon entre 2014 y 2017 en el marco del Shanghai Diabetic Complication Screening Project (SDCSP), el Nicheng Diabetes Screening Project (NDSP), y el programa Nationwide Screening for Complications of Diabetes (NSCD), en China. En particular, se proveen 4 imágenes por paciente, 2 por cada ojo, tomadas con centro en la mácula y en el disco óptico. El conjunto también incluye particiones predeterminadas de entrenamiento, validación y test. Sin embargo, dado que las etiquetas de grado de RD según la ETDRS son de acceso público únicamente para los estudios de entrenamiento y validación, se decidió reservar el primer conjunto para entrenamiento y emplear el segundo para test, descartando los de test provistos originalmente.

4.1.3. APTOS2019

APTOS2019⁵ contiene imágenes adquiridas en el Hospital de Ojos Aravind de la India, obtenidas bajo diferentes condiciones de captura y con distintas cámaras en el marco de una iniciativa regional de tamizado poblacional. Las imágenes cuentan con cierto nivel de ruido tanto a nivel de las capturas (es decir, errores de adquisición, mala calidad, etc.) como de las etiquetas de grado de RD asociadas (asignaciones de categorías diagnósticas erróneas). Dado que las redes neuronales son capaces de lidiar con este tipo de ruido durante el entrenamiento incluso para su propio beneficio [18], se incluyó la partición de entrenamiento predeterminada en este conjunto como parte de la finalmente empleada para el aprendizaje de la red. Ningún estudio de APTOS2019 se reservó para evaluación para evitar potenciales sesgos.

⁴ <https://isbi.deepdr.org/data.html>

⁵ <https://www.kaggle.com/c/aptos2019-blindness-detection/data>

4.1.4. EyePACS

EyePACS⁶ (también conocido como Kaggle en la literatura) es el mayor conjunto de FFO con etiquetas de grado de RD liberado públicamente para el entrenamiento y validación de algoritmos de IA para reconocimiento de RD. Las imágenes, correspondientes al ojo izquierdo y derecho de pacientes de Estados Unidos, presentan resoluciones variables (desde 320 x 211 píxeles a 5.184 x 3.456 píxeles), fueron tomadas con cámaras diversas y provienen de los registros de la plataforma de tamizado EyePACS. A pesar de que muchas presentan inconvenientes de calidad [19], se respetaron las particiones de entrenamiento y test predeterminadas en el conjunto, extrayendo un 10% de los pacientes de entrenamiento para integrar el conjunto de validación, utilizando el 90% restante para entrenamiento, y conservando el de test para la evaluación final del algoritmo.

4.1.5. IDRiD

IDRID [20] brinda 516 FFO midriáticas centradas en la mácula, capturadas con una cámara Kowa VX-10a de 50° de campo visual (FOV) entre 2009 y 2017 en la Clínica de Ojos de Nanded, India, todas a partir de pacientes diabéticos. Los estudios tienen una resolución de 4.288 x 2.848 píxeles, y están ya separados en particiones predeterminadas de entrenamiento y test pero no de validación, por lo que se separaron un 10% de las de entrenamiento para construir este último conjunto, respetando el resto de divisiones establecidas.

4.1.6. FCM-UNA

Denominamos FCM-UMA al conjunto de 757 FFO midriáticas liberadas en [21] para evaluación de algoritmos de clasificación de RD. Las imágenes, todas centradas en

⁶ <https://www.kaggle.com/c/diabetic-retinopathy-detection/data>

la mácula, fueron adquiridas con una cámara Zeiss Visucam 500 de 45° de FOV, a una resolución de 2.124 x 2.056 píxeles, en el Departamento de Oftalmología del Hospital de Clínicas de la Facultad de Ciencias Médicas (FCM) de la Universidad Nacional de Asunción (UNA), Paraguay, y categorizadas en 7 clases (1 - sin signos de RD, 2 - RD no proliferativa leve o temprana, 3 - RD no proliferativa moderada, 4 - RD no proliferativa severa, 5 - RD no proliferativa muy severa, 6 - RD proliferativa, 7 - RD proliferativa muy avanzada). Para construir la etiqueta de referibilidad, se agruparon las clases 1 y 2 como no referibles, y las restantes como RD referible. Todas estas imágenes se reservaron para evaluar la efectividad del algoritmo.

4.1.7. 1000Fundus

1000Fundus [22] contiene 1.000 imágenes obtenidas en el Joint Shantou International Eye Center en China, capturadas entre 2009 y 2018 utilizando cámaras Zeiss FF450 Plus IR y Topcon TRC-50DX midriáticas, con resoluciones variables. Cada imagen tiene asociada una etiqueta indicando la presencia/ausencia de 39 observaciones diferentes, que incluyen enfermedades, defectos de captura o signos de relevancia clínica. Los casos de RD están agrupados en 3 categorías (1 - leve, 2 - moderada, 3 - avanzada), por lo que la categoría 1 y los casos sanos se seleccionaron como no referibles y las restantes como RD referible. El conjunto resultante de 144 imágenes se utilizó como conjunto de test, mientras que la totalidad de las 1.000 imágenes se empleó separadamente para analizar la respuesta del algoritmo ante la presencia de distintas observaciones (Sección 5.5).

4.1.8. DR2

DR2 [23] contiene 435 imágenes de personas diabéticas capturadas con una cámara Topcon TRC-NW8 no midriática en el Departamento de Oftalmología de la

Universidad Federal de São Paulo en Brasil. Todas las FFO se encuentran reducidas a una resolución de 867 x 575 píxeles y están etiquetadas como referibles o no por parte de al menos 2 expertos, quienes determinaron la referibilidad en función de observaciones que consideraron relevantes en la imagen, y no para un tipo particular de lesión. Por este motivo, el conjunto fue utilizado en su totalidad para los datos de test, ya que permite estudiar la performance del algoritmo ante un criterio diagnóstico alternativo.

4.1.9. MESSIDOR 2

MESSIDOR-2 [24] contiene 1.748 imágenes de los ojos izquierdo y derecho de 874 pacientes, obtenidas de tres departamentos oftalmológicos de Francia utilizando una cámara Topcon TRC NW6. 800 estudios fueron adquiridos con midriasis, y los restantes sin. Cada imagen cuenta con un grado de RD asociado entre 0 (sin signos) y 3 (RD proliferativa), producido según el número de lesiones presentes [24]. En línea con lo observado en la literatura [23, 25], se agruparon las categorías 0 y 1 como sin RD referible, y las restantes como con RD referible. 2 pacientes (4 imágenes) fueron descartados por no contar con anotaciones.

4.1.11. DIARETDB1

DIARETDB1 [26] cuenta con 89 imágenes, 84 de las cuales presentan signos de RD leve o preproliferativa y 5 consideradas normales, capturadas utilizando una cámara Zeiss FF 450^{plus} en el Hospital Universitario de Kuopio en Finlandia. En lugar de brindar anotaciones describiendo el grado de la enfermedad, se proveen segmentaciones manuales de lesiones típicas de la RD (exudados, microaneurismas y hemorragias) generadas por varios expertos. Para producir las etiquetas de referibilidad, se aplicó la tabla de conteo de lesiones documentada en [24],

originalmente empleada para etiquetar MESSIDOR 2, y luego se agruparon las categorías 0 y 1 como sin RD referible y 2 y 3 como RD referible. Todas las imágenes se incluyeron en el conjunto de test.

4.1.12. Martínez y HEC

Para la evaluación sobre datos de pacientes nacionales, se realizó una recolección retrospectiva de imágenes a partir de las bases de datos de dos establecimientos de salud argentinos, el Centro Oftalmológico Martínez de Pehuajó (conjunto Martínez) y el Hospital de Alta Complejidad En Red “El Cruce” Dr. Néstor Carlos Kirchner de Florencio Varela (conjunto HEC), ambos ubicados en la provincia de Buenos Aires. El protocolo para trabajar con estos estudios fue avalado oportunamente por el Comité de Ética del Hospital El Cruce (dictamen 25/2020). El conjunto Martínez está constituido por 484 imágenes obtenidas con un dispositivo Cristal Vue NFC-700, no midriático y con 45° de FOV, 30 de las cuales presentan signos de referibilidad de RD. Nótese que imágenes capturadas con este modelo de cámara no fueron empleadas para el entrenamiento del algoritmo. Las del conjunto HEC, por otro lado, corresponden a un atlas de 35 estudios constituido en el hospital para la formación y discusión con los profesionales del Servicio de Oftalmología. Las imágenes fueron obtenidas con una cámara retinal no midriática Topcon TRC-NW8, similar a la utilizada en el conjunto DR2. Los estudios de ambos conjuntos fueron etiquetados manualmente, cada una por una oftalmóloga diferente, indicando tanto si se trataba de casos de RD referible como otras observaciones diagnósticas.

4.2. Preprocesamiento de las imágenes

En los ejemplos ilustrativos de la Figura 1 puede observarse una gran variabilidad en las imágenes, sobre todo en resolución e iluminación. Las redes neuronales requieren para su entrenamiento lotes de imágenes de entrada con una resolución

uniforme, por lo que es necesario cambiar el tamaño de todas ellas a una dimensión única. En la Figura 2 se presenta una FFO de fondo de ojo de ejemplo y el resultado obtenido al reescalarla a una resolución cuadrada. Puede observarse que ocurre una degradación significativa en la relación de aspecto original al aplicar esta transformación. Esto se debe a que, dependiendo del dispositivo de captura, a menudo puede ocurrir que grandes porciones de la imagen carezcan de información de la retina, por hallarse fuera de la región del FOV de la cámara. En consecuencia, reducir la resolución original de la imagen (en el ejemplo, 3.168 x 4.752 píxeles) a una alternativa cuadrada (512 x 512 píxeles) alterará su contenido y disminuirá la cantidad de píxeles con información de la retina.

Para resolver este inconveniente, se desarrolló una estrategia de preprocesamiento basada en dos etapas, una primera enfocada en recortar el estudio en el área del FOV para mantener la información de la retina al realizar el reescalado, y una segunda que procura incrementar el contraste en escenarios de iluminación despareja. Ambas etapas se ilustran para la misma imagen de ejemplo en la segunda parte de la Figura 2.

Para recortar la imagen en torno al FOV, es necesario primero estimar su área de manera automática, generando una máscara binaria que determine qué píxeles están ubicados dentro del área de interés de la imagen. Para tal fin, el algoritmo propuesto genera inicialmente una matriz del mismo tamaño de la imagen de entrada, donde cada píxel contiene la suma de las intensidades de color rojo, verde y azul de la imagen de entrada. Estos valores son luego umbralizados utilizando el método automático de Otsu [27] para obtener una primera máscara binaria. Debido a que en casos de imágenes de mala calidad puede ocurrir que la región resultante presente falsos negativos dentro del área del FOV, se corrige esta máscara inicial usando una operación de eliminación de agujeros basada en morfología matemática. En la Figura 2 (c) se muestra un ejemplo de la estimación inicial del FOV a partir de la imagen de entrada, y un detalle de la misma. Obsérvese que la máscara, a pesar

de tener falsos negativos en el área sobreexpuesta y no cerrar completamente en la parte superior, incluye parte del borde circular del FOV del lado izquierdo (indicado con la flecha). Esto permite que puedan identificarse sobre ella las coordenadas del rectángulo que mejor la envuelve, y emplearlas luego para recortar la imagen original. La Figura 2 (d) muestra un ejemplo de la imagen final, recortada y reescalada automáticamente a 512 x 512 píxeles. Obsérvese que la relación de aspecto en esta imagen se conserva mejor que en la Figura 2 (b).

Para mejorar el contraste de la imagen en situaciones de mala iluminación o con iluminación parcial, se aplicó además un método de preprocesamiento estándar utilizado con anterioridad en [25]. Éste requiere contar con una mejor estimación del FOV que la utilizada para el recorte, ya que es necesario realizar cálculos en el interior del mismo. Por ello se corrige inicialmente el FOV calculando la cápsula convexa de la detección primaria (Figura 2 (e)), obteniendo una estructura compacta y sin huecos incluso en casos en los que la máscara inicial es cóncava (como en el ejemplo). Luego, se estima el fondo de iluminación de la imagen a partir de la aplicación de un filtro gaussiano con desvío igual a 1/90 del tamaño horizontal de la imagen recortada (Figura 2 (f)). Para evitar efectos indeseados en los bordes del FOV, este cálculo se realiza en una versión alternativa de la imagen original en la que se asignan a los píxeles localizados fuera de esta región los valores promedio de rojo, verde y azul calculados en el interior del FOV. Finalmente, este fondo B es utilizado para obtener una nueva imagen $I_{mejorada}$ aplicando la fórmula:

$$I_{mejorada} = 4 \times I_{original} - 4 \times B + 128$$

donde $I_{original}$ es la imagen de entrada original, recortada. El resultado es una imagen como la de la Figura 2 (g), que presenta una mejor definición de las estructuras retinianas en las zonas de iluminación despareja (ver detalle en Figura 2 (h)). Además, esta estrategia permite incrementar el contraste de algunas lesiones

características de la RD, como en el caso de las microaneurismas y hemorragias señaladas por las flechas en la Figura 2 (h).

4.3. Modelo de inteligencia artificial desarrollado

Las imágenes de entrenamiento, preprocesadas con la estrategia descrita en la Sección 4.2, fueron utilizadas para el aprendizaje de nuestro modelo de IA para el reconocimiento automático de la RD referible. El algoritmo desarrollado corresponde a una red neuronal convolucional profunda con arquitectura residual, conocida en la literatura como ResNet [28], preentrenada sobre datos de ImageNet. El detalle de la arquitectura se presenta en la Figura 3, omitiendo las capas de normalización por lotes (o *Batch Normalization*, BN) y las unidades de rectificación no lineales (ReLUs) por cuestiones de espacio. La red utilizada cuenta con un total de 18 capas principales, agrupadas en 6 subconjuntos. La entrada es una imagen a color de 512 x 512 píxeles, normalizada previamente con intensidades en el intervalo [-1, +1]. Sobre ella se aplica inicialmente una capa convolucional con 64 filtros de 7 x 7 píxeles y un desplazamiento de 2 píxeles, seguida de una operación de BN y una ReLU. La resolución de las activaciones de salida de este subconjunto se reduce a la mitad mediante una capa de *max pooling* con núcleos de tamaño 2 x 2 y desplazamientos de 2 píxeles. Su correspondiente salida, de dimensiones 256 x 256 píxeles, se utiliza como entrada para un subconjunto de 4 capas convolucionales, todas ellas con 64 filtros de 3 x 3 píxeles cada una, seguidas cada una por una capa de BN y una ReLU. Obsérvese que cada 2 capas se utiliza una conexión por adelantamiento, típica de las ResNets, que concatena los mapas de activaciones de la última capa con la entrada al bloque y permite el entrenamiento eficaz de este tipo de redes profundas [28]. Esta secuencia de trabajo con 4 convoluciones es repetida secuencialmente en los 3 grupos de capas posteriores, con la sola modificación de que la primera capa convolucional de cada bloque incrementa al doble el número de filtros que utilizarán las capas siguientes, y que utiliza un desplazamiento de 2

píxeles en lugar de 1 para reducir la dimensionalidad de la entrada en lugar de una capa de *max pooling*. Las conexiones de adelantamiento ubicadas entre dos bloques sucesivos con dimensiones diferentes (indicadas con líneas punteadas en la Figura 3) utilizan rellenos con ceros para asegurar la validez de la concatenación. Tras la sucesión de los 4 grupos de 4 convoluciones cada uno, se utiliza una operación de *average pooling* adaptativo para reducir la dimensionalidad del tensor de salida de la última convolución (de 16×16 píxeles x 512 características) a una representación vectorial de 512, que se utilizan como entrada para una capa totalmente conectada (*fully connected*) que realiza la clasificación en las categorías “sin RD referible” y “con RD referible”. Sobre los dos *logits* de salida del modelo se aplica una función de activación *softmax*, que permite transformarlos en probabilidades para cada clase. Así, la salida del modelo son valores de probabilidad mutuamente excluyentes que indican qué tan factible es que la imagen de entrada corresponda a un caso de RD referible o a uno sin RD referible ($1 - \text{probabilidad de RD referible}$). Cuando una probabilidad se aproxima más a 1 que la otra, el modelo tiene mayor certeza respecto a su respuesta, mientras que cuando ambas probabilidades son similares (en torno al 50%), el modelo presenta incertezza.

El algoritmo fue implementado en el lenguaje de programación Python versión 3.9 utilizando la librería para aprendizaje profundo PyTorch versión 1.10 [29]. El entrenamiento del modelo se realizó por un máximo de 150 épocas, deteniéndolo tempranamente si el área bajo la curva ROC (AUC) sobre el conjunto de validación no mejoraba por un máximo de 40 épocas. Se utilizó la entropía cruzada como función de pérdida, y el algoritmo de optimización Adam con sus parámetros por defecto para la minimización de la misma. Se empleó un tamaño de lote de 128 imágenes por iteración y una tasa de aprendizaje inicial de 0.0001. Para favorecer el alcance de mejores mínimos, la tasa de aprendizaje se redujo a la mitad cada vez que el AUC sobre los datos de validación no lograba mejorar en 20 épocas. Además, para reducir el riesgo de sobreentrenamiento se utilizó decaimiento de pesos con una

tasa de 0.001 y se aplicó una estrategia de aumentación de imágenes. Esta última consistió en la aplicación sucesiva, aleatoria y en línea de las transformaciones indicadas en la Tabla 2, que permiten obtener imágenes alteradas artificialmente y simular la existencia de un mayor número de estudios. Las transformaciones corresponden a las implementaciones provistas en el paquete torchvision versión 0.11 [29], y sus magnitudes fueron seleccionadas experimentalmente evaluando la mejora de los resultados sobre los datos de validación. Todas ellas fueron aplicadas sobre las imágenes recortadas y reescaladas con la estrategia de preprocesamiento descrita en la sección 4.2, pero antes de la aplicación de la mejora de contraste, para reducir la influencia de esta última sobre la transformación. La Figura 4 presenta un ejemplo de la aplicación sucesiva de las transformaciones sobre una imagen.

El modelo fue entrenado utilizando una instancia en línea en la plataforma de ciencia de datos en la nube Saturn Cloud (Saturn Cloud, Inc., Nueva York, EE.UU.), con un procesador de 8 núcleos, 61 GB de memoria RAM, 300 GB de almacenamiento en disco y 1 unidad de procesamiento gráfico NVIDIA V100 con 16 GB de memoria. Cada entrenamiento en esta plataforma demoró aproximadamente unas 12 horas. Para la evaluación, el modelo fue ejecutado en una computadora de escritorio con un procesador Intel i7 de 10ma generación, 32 GB de memoria RAM, 1 TB de almacenamiento en disco y 1 unidad de procesamiento gráfico NVIDIA 3060 con 12 GB de memoria. El tiempo de ejecución promedio por imagen para los estudios del conjunto Martínez (de resolución 4096 x 3072 píxeles) en este equipo, incluyendo las operaciones de preprocesamiento, fue de 1.24 segundos.

Para el estudio cualitativo de los resultados del modelo se extrajeron mapas de activación de clases mediante la técnica XGrad-Cam [30], empleando la implementación de la librería Torchcam [31] sobre los bloques de capas convolucionales 2, 3 y 4 del modelo, combinándolas en un único mapa. Se extrajo en cada caso el mapa correspondiente a la clase predicha (es decir, si el modelo predijo que el caso era no referible, se obtuvo el mapa de atribuciones para esa clase, y

viceversa), para identificar qué regiones de la imagen fueron tenidas en cuenta por el modelo para brindar su respuesta. Dado que el mismo se calcula sobre la imagen recortada y reescalada, una vez obtenido fue reescalado a la resolución de la imagen original para favorecer la comparativa con la misma. Para su representación gráfica se utilizó un mapa de calor, donde las regiones con tonalidades rojas se asocian a activaciones altas (regiones tenidas en cuenta) y las azules a activaciones bajas (regiones que el modelo no tuvo en cuenta).

4.4. Análisis de datos

El método propuesto fue evaluado cuantitativamente mediante métricas estándar de clasificación binaria, incluyendo área bajo curva ROC (AUC), sensibilidad (Se) y especificidad (Esp). El AUC se obtuvo a partir de las estimaciones de probabilidad de RD referible realizadas por el modelo, utilizando la implementación provista en el paquete Scikit Learn [32]. La sensibilidad y especificidad, por su parte, se calcularon con código propio, tomando como casos con RD referible a aquellos cuya probabilidad asociada, según la predicción del modelo, sea igual o mayor a 0.5. En todos los casos se incluyen los intervalos de confianza (IC) del 95%. Para obtener los IC del AUC se aplicó la técnica de *bootstrap* con reposición para n = 1.000 muestras. Para los de la sensibilidad y especificidad se utilizó una implementación Python⁷ que aplica el método de Wilson [33].

5. Resultados

En las siguientes subsecciones se presenta el análisis racional de cada uno de los experimentos realizados y los resultados obtenidos. En la sección 6 se discuten los mismos y sus implicancias clínicas.

⁷ <https://gist.github.com/maidens/29939b3383a5e57935491303cf0d8e0b>

5.1. Evaluación de las predicciones a nivel imagen

Para ser aplicado con éxito para tamizar casos que requieran ser estudiados por un experto, el modelo propuesto debe ser capaz de identificar con gran precisión si una imagen dada corresponde o no a un paciente con RD referible, independientemente del dispositivo de captura empleado. Para simular este escenario, se cuantificaron los valores de AUC, sensibilidad y especificidad y sus respectivos IC 95% en cada uno de los conjuntos de test planteados en la sección 4.1. La Tabla 3 resume los resultados obtenidos. Para facilitar la comparación con el estado del arte, se incluyen los resultados reportados por los trabajos más recientes de la literatura sobre algunos de los conjuntos de test. Nótese que no todos los trabajos reportan sobre todos los grupos de imágenes aplicados en este trabajo.

En la Figura 8 se presentan múltiples ejemplos de imágenes de entrada, predicciones realizadas por el algoritmo y sus respectivos mapas de atribución. En los casos en los que el algoritmo determina no referibilidad, se observa una tendencia a presentar mayores activaciones en la región de la mácula y las arcadas vasculares principales (ver primera imagen del ojo izquierdo en la Figura 8 (a) y primeras dos imágenes de la Figura 8 (b)). En los casos predichos como referibles, las activaciones se concentran mayoritariamente sobre lesiones hemorrágicas, exudaciones y algodonosas presentes en la mácula, el disco óptico o la periferia.

5.2. Evaluación utilizando múltiples imágenes por paciente

En la práctica clínica tradicional es normal que el oftalmólogo produzca múltiples imágenes de cada ojo del paciente, o que analice la retina mediante una oftalmoscopía recorriendo diversas áreas en busca de signos de preocupación. En un enfoque telemédico, es normal que se capturen múltiples estudios del paciente (por lo general al menos uno por cada ojo [40]) y que todos sean transmitidos al centro de informe. Un paciente dado es considerado entonces referible si presenta signos de RD preocupantes en al menos uno de sus ojos, o al menos una de las

imágenes. Para evaluar el comportamiento al contar con al menos dos estudios por paciente, se utilizaron los datos de MESSIDOR 2 y EyePACS (que proveen una imagen de cada ojo para cada paciente) y DeepDRID (que brinda dos imágenes por cada ojo de cada paciente). En cada conjunto se produjeron etiquetas a nivel de paciente asumiéndolo como referible si alguna de sus imágenes era etiquetada como tal. Este criterio se aplicó tanto sobre las etiquetas manuales como sobre las predicciones del algoritmo para poder compararlas entre sí. Para el cálculo de la probabilidad de RD referible del paciente, se tomó el máximo de las probabilidades asignadas a sus imágenes. Los resultados en los tres conjuntos se presentan en la Tabla 4. Se incluyen también los resultados de dos trabajos de la literatura que también realizaron este experimento sobre MESSIDOR 2 y EyePACS.

En la Figura 8 (a) se ejemplifican resultados cualitativos obtenidos sobre las dos imágenes del ojo izquierdo y derecho de un paciente del conjunto DeepDRID sin signos de RD referible en el ojo izquierdo, pero con RDP en el derecho. Aunque el modelo identifica no referibilidad en el primero de los estudios del ojo izquierdo, predice referibilidad con un valor alto de probabilidad en la segunda imagen de ese ojo y en ambas del ojo derecho, por lo que lo identifica como referible.

5.3. Evaluación según calidad de la imagen

En el contexto de una red de teleoftalmología, es de esperar que los técnicos que capturan los estudios puedan cometer errores y produzcan algunas imágenes de mala calidad que afecten el diagnóstico posterior. Para estudiar la influencia de la calidad del estudio sobre la respuesta del método propuesto, se realizaron evaluaciones adicionales diferenciando distintos subgrupos de estudios, según su calidad. Para este análisis fue necesario recurrir a etiquetas de calidad para cada imagen estudiada, únicamente disponibles para algunos estudios de las bases de datos de test de EyePACS y DeepDRID. Las curvas ROC obtenidas para cada subgrupo de calidad se presentan en la Figura 5, acompañadas de ejemplos de cada

tipo de calidad de imagen. En la Figura 5 (a) se observan las curvas obtenidas sobre un subconjunto de 17.515 imágenes del conjunto de test de EyePACS, etiquetados en [19] según 3 categorías: rechazables (estudios que no podrían emplearse jamás para diagnosticar RD, N = 9.111), usables (que presentan dificultades de captura y artefactos pero permiten observar estructuras anatómicas de interés, N = 4.898) y buenas (de calidad suficiente para diagnóstico, N = 3.506). En la Figura 5 (b) se utilizan 46.028 imágenes del conjunto de test de EyePACS, pero etiquetadas según [41] en imágenes de mala calidad (imposibles de usar para diagnóstico, N = 944) y buena calidad (usables para diagnóstico, tengan o no artefactos, N = 45.084). Finalmente, la Figura 5 (c) incluye los resultados sobre las 400 imágenes del conjunto de validación de DeepDRID, utilizado en este trabajo como de test, donde los estudios se clasifican en calidad buena (suficiente para diagnóstico, N = 182) y malas (suficiente para diagnóstico pero con artefactos que ameritan recaptura, N = 218). En la Figura 8 (a) se ejemplifican algunos resultados obtenidos por el algoritmo propuesto sobre imágenes del conjunto DeepDRID señaladas en el mismo como de mala calidad. En la segunda imagen del ojo izquierdo, donde se predice referibilidad, el mapa de activación asociado exhibe respuestas en torno a artefactos brillantes producto de una captura errónea, identificándolos como potenciales lesiones. En la primera imagen del ojo derecho, por el contrario, el algoritmo reacciona a la región entre los dos vasos sanguíneos principales de la parte superior y a la región ubicada inmediatamente debajo del nervio óptico, mientras que en la segunda se centra sobre todo en la región macular, donde se observan algunos exudados.

5.4. Reconocimiento de casos de gravedad variable

La RD no tratada evoluciona naturalmente en complicaciones de gravedad creciente, clasificadas según la ETDRS en 5 grados (ver sección 4.1). El propósito del algoritmo desarrollado es reconocer con gran precisión aquellos casos que requieran ser revisados por un profesional. No obstante, es importante asegurar una tasa baja de

falsos negativos en los que corresponden a condiciones graves que requieren tratamiento urgente (es decir, RDNP severa o RD proliferativa). Para evaluar las tasas de falsos negativos/positivos en cada grado avanzado de la enfermedad, se cuantificó la performance del algoritmo para reconocer casos de RD no referible vs. RDNP moderada, primero, luego RD no referible vs. RDNP severa, y finalmente RD no referible de RD proliferativa (RDP). Para tal fin se utilizaron subconjuntos extraídos de DDR, IDRiD, MESSIDOR 2, EyePACS, DeepDRiD y FCM-UMA, ya que estos proveen etiquetas indicando el grado de la enfermedad. La Figura 6 presenta las curvas ROC y los valores de AUC obtenidos para diferenciar cada grado respecto a sujetos sin RD referible. La Tabla 5 incluye además los valores de sensibilidad y especificidad, junto a sus IC. En la Figura 8 (b) se presentan algunos resultados cualitativos obtenidos sobre imágenes del conjunto FCM-UNA con diferentes grados de RD. En los casos indicados como no referibles (sin signos de RD o con RDNP leve), la atención del modelo se concentra mayormente en la mácula y las arcadas vasculares. En el caso de RDNP moderada se observa una predicción de referibilidad con una certeza del 79.5%, y que las activaciones se concentran en la pequeña hemorragia ubicada en la región superior del disco óptico y en una lesión dentro de la papila. En el caso con RDNP severa, se presentan activaciones en los exudados cercanos a la mácula y en algunas hemorragias, mientras que en el señalado como con RDNP muy severa la atribución principal se localiza en una región sin lesiones aparentes, y otras menos fuertes sobre múltiples hemorragias pequeñas cercanas a la fóvea. En el caso con RDP, finalmente, no se observan activaciones en la zona con neovascularizaciones, sino sobre algunas lesiones rojas.

5.5. Relación entre las salidas y otras observaciones en la imagen de entrada

El método propuesto produce como salida un valor de probabilidad que determina cuál es el riesgo de que la imagen analizada corresponda a un caso de RD referible. Para estudiar la sensibilidad de esta respuesta a las características de la imagen de

entrada, se realizó un análisis cuantitativo utilizando los 1000 estudios de la base de datos 1000Fundus, que provee imágenes categorizadas según 39 observaciones diferentes. Las mismas corresponden tanto a diagnósticos específicos (por ejemplo, miopía patológica, oclusión de rama venosa retiniana, etc.) como a características anatómicas del paciente (ej. copa óptica grande) u observaciones sobre la imagen (ej. fondo difuminado, presencia de lesiones por láser, etc.), agrupadas además según se trate de situaciones no referibles o referibles a un profesional o de escenarios de urgencia. El análisis, resumido en el boxplot de la Figura 7, consistió en estudiar la distribución de las probabilidades de RD referible obtenidas sobre cada subconjunto de imágenes, categorizadas según cada observación. Las muestras ubicadas más a la derecha (es decir, por encima del valor 0.5) son identificadas por el algoritmo como potenciales casos de RD referible, y las ubicadas a la derecha como sin RD referible. Se observa que las probabilidades de RD referible más altas se obtuvieron en casos con RDNP moderada y severa o RDP, pero también en casos observados como padecientes de oclusión de rama venosa retiniana o de la vena central o con retinopatía hipertensiva severa, o que presentan partículas en el vítreo, exudados duros masivos o lesiones algodonosas, o en casos en los que el fondo de ojo se observa difuminado pero se sospecha presencia de signos de RDP. En la Figura 8 (c) se incluyen algunos casos de ejemplo y sus mapas de activación. En el paciente con marcas de láser, la red neuronal predijo referibilidad e indicó como potenciales lesiones a algunas de esas marcas. En el de oclusión de rama venosa retiniana, se realiza una predicción sin mucha certeza (53.9% de referibilidad), donde las activaciones se localizan encima de exudaciones y hemorragias. El caso con lesiones algodonosas se identifica como RD referible, y las atribuciones se localizan sobre las mismas. Algo similar ocurre en el caso de retinopatía hipertensiva severa. En el que presenta partículas en el vítreo, el mapa presenta activaciones sobre las mismas, probablemente confundiéndolas con exudados. Finalmente, en el caso con

RDP y el fondo difuminado se observan activaciones sobre algunas de las hemorragias ubicadas sobre y debajo del disco óptico.

5.6. Evaluación sobre imágenes de pacientes argentinos

Para estudiar si efectivamente el método propuesto presenta la robustez suficiente como para ser aplicado en un escenario clínico local, se realizó una evaluación cuantitativa y cualitativa sobre los bancos de datos Martínez y HEC. Los resultados obtenidos se presentan al final de la Tabla 3, en términos de AUC, sensibilidad y especificidad, y las curvas ROC en cada conjunto se ilustran en la Figura 9.

En la Figura 9 (a) se presentan algunos casos de ejemplo del conjunto Martínez. El primero corresponde a un caso no referible con un mal enfoque del polo posterior, que el algoritmo detecta correctamente como no referible. La mayoría de las activaciones se localizan sobre algunos de los vasos principales y en el área macular. La imagen de su derecha pertenece a un paciente con RD referible, que presenta exudaciones y lesiones hemorrágicas localizadas en torno a la mácula. El modelo identifica esta zona como de riesgo en el mapa de activaciones, y lo señala como un caso referible. En la segunda fila de ejemplos se observan dos resultados no coincidentes con el valor esperado. El de la izquierda es un paciente con cataratas cuya imagen no es nítida y presenta artefactos producto de suciedad en el lente del retinógrafo. A pesar de que no corresponde a un caso con RD referible, es detectado por el modelo como referible con un valor de probabilidad incierto (cercano al 50%), basando su respuesta en los defectos de captura ubicados en el borde del FOV. El caso de la derecha, por otro lado, pertenece a un paciente con maculopatía miópica que es detectado por el algoritmo como un potencial caso de RD referible, basándose sobre todo en las anomalías presentes en la región foveal.

Los ejemplos de la Figura 9 (b) presentan algunos casos del conjunto HEC. El primero de ellos corresponde a un sujeto sin lesiones compatibles con RD, al que se le sugiere un control clínico por presentar cambios compatibles con hipertensión

arterial (en particular alteraciones de realización arteriovenosa y cruces arteriovenosos). Nótese que el algoritmo lo reconoce como no referible, y presenta activaciones sobre uno de estos cruces. El segundo ejemplo ubicado a su derecha corresponde a un paciente con RD referible, que exhibe exudados duros en la región de la mácula, exudados blandos y duros en la arcada temporal superior y hemorragias aisladas. Tanto los exudados duros del área macular como los blandos fueron identificados por el algoritmo en su mapa de activaciones, clasificando al paciente como referible. Los ejemplos de la parte inferior de la Figura 9 (b), por otro lado, se asocian a casos detectados erróneamente. El de la derecha no corresponde a un paciente con RD referible, pero sí con signos compatibles con obstrucción arterial no aguda, sin hemorragias pero con exudados. El algoritmo lo identifica como un caso potencial de RD referible, detectando algunos de ellos como signos de la enfermedad. El otro ejemplo es el de un paciente sin RD referible pero con lesiones de fotocoagulación láser por fuera de las arcadas vasculares, cuya imagen está fuera de foco e impide analizar en detalle la zona macular. El algoritmo lo reconoce como un caso referible, e indica que tuvo en cuenta algunas de las lesiones por el tratamiento láser.

6. Discusión

6.1. Evaluación general de los resultados

Según se observa en la Tabla 3, el modelo propuesto obtiene valores de AUC superiores a 0.94 para el reconocimiento de RD referible en todos los conjuntos de test empleados, lo que resalta su robustez a cambios en el dispositivo de captura o la población de personas diabéticas estudiadas. La sensibilidad se ubica por encima del 73% (llegando en el caso de 1000Fundus y el HEC a alcanzar el 100%), mientras que la especificidad se mantiene superior al 86%. Es importante señalar que estos valores fueron calculados sin optimizar el punto operativo de la curva ROC como en

otros trabajos [5], sino que tomando como referibles a aquellos con probabilidad mayor o igual a 0.5. La exploración de otros potenciales umbrales permitiría ajustar la sensibilidad y la especificidad a los efectos de, por ejemplo, reducir el número de falsos negativos (que constituyen casos de riesgo que deberían ser presentados sí o sí al oftalmólogo) a costa de una tasa de falsos positivos mayor (que elevaría el costo de análisis en el nodo de informes, ya que requeriría informar estudios sin RD referible). No obstante, del análisis realizado acerca de la habilidad del modelo para diferenciar grados avanzados de la enfermedad (Sección 5.4, Tabla 5 y Figura 6) se desprende que es capaz de localizar con gran precisión los casos con RDNP severa y RDP, que son los más graves y que requieren tratamiento más urgente.

La evaluación de la Sección 5.2 permite comprobar cómo el uso de múltiples imágenes por paciente puede favorecer el hallazgo de signos de la RD que puedan ser ignorados en un único estudio. Si se comparan los resultados presentados en la Tabla 4 con los de la Tabla 3, para los cuales se usó una sola imagen por paciente, es posible observar una mejora en la sensibilidad a costa de una disminución en la especificidad. En MESSIDOR 2, por ejemplo, la sensibilidad aumenta levemente de 0.895 a 0.907, pero la especificidad disminuye de 0.941 a 0.920, lo que produce una caída leve del AUC de 0.970 a 0.967. Un comportamiento similar se observa en EyePACS y en DeepDRID. En esta última, utilizar un par de imágenes por cada ojo del paciente (es decir, 4 en total) permite incrementar la sensibilidad casi en un 10% (de 0.885 a 0.980) a costa de una reducción en la especificidad (0.868 vs. 0.800). Este detalle cobra relevancia teniendo en cuenta que un sistema autónomo para tamizado de casos de riesgo debe presentar un número acotado de falsos positivos pero sobre todo reducir al mínimo posible los falsos negativos. Incorporar un mayor número de imágenes por paciente puede entonces impactar positivamente en el sistema, mejorando su robustez para reconocer casos positivos al costo de realizar algunas detecciones adicionales incorrectas. No obstante, es importante tener en cuenta que el criterio empleado para integrar las respuestas obtenidas para cada ojo

consistió en nuestro caso simplemente en tomar la probabilidad máxima de referibilidad para cada paciente, algo que a priori puede conducir a un mayor número de falsos positivos. En la Figura 8 (a) puede verse un ejemplo claro de este comportamiento, ya que el algoritmo señala al ojo izquierdo como referible siendo que no se identificaron signos en la primera imagen pero sí en la segunda (que, además, son artefactos de captura), lo que en términos de probabilidad máxima acaba implicando una predicción de referibilidad. En [39], los autores sugieren entrenar un clasificador por separado que tome características obtenidas para cada imagen y las combine para generar una única respuesta. Aunque aplicar dicha estrategia en nuestros datos queda fuera del alcance de este artículo, explorar otras formas alternativas para combinar respuestas de varias imágenes de un mismo paciente puede ayudar a mejorar los resultados obtenidos, eventualmente incrementando la sensibilidad sin ocasionar caídas bruscas de la especificidad.

Otro punto importante tiene que ver con la calidad del estudio de entrada. En imágenes con calidad aceptable o con artefactos que no impiden observar las estructuras principales de la retina, el algoritmo es capaz de obtener resultados muy similares a los que puede lograr sobre estudios óptimos. Esto se observa en la Figuras 5 (a) y (c), donde el método evaluado en las imágenes “Usables” y “Buenas” de EyePACS y en las “Malas” y “Buenas” de DeepDRID, respectivamente, reporta curvas ROC muy similares entre sí. En los resultados cualitativos presentados en la Figura 8 (a), por otro lado, es posible comprobar que el algoritmo puede reconocer signos de la RD incluso cuando la imagen presenta defectos de iluminación o una definición subóptima de las estructuras retinianas. No obstante, en casos donde las imágenes son de muy mala calidad se observa una notoria desmejoría en la respuesta del algoritmo, que produce una reducción importante en su AUC (Figuras 5 (a) y (b)). Esto puede deberse a problemáticas como la representada en la segunda imagen del ojo izquierdo en la Figura 8 (a), donde el modelo reconoce artefactos de

captura como potenciales lesiones de la RD. Distinguir de manera anticipada los casos de mala calidad (ver Sección 6.2) podría aliviar este inconveniente.

En lo que refiere a la comparativa con otros trabajos de la literatura, disponible en las Tablas 3 y 4, se observa que el método propuesto es capaz de superar en numerosos escenarios la performance de modelos más complejos. En el conjunto DDR, por ejemplo, nuestro algoritmo supera en AUC al de detección de lesiones propuesto en [34]. Algo similar ocurre en el conjunto de test de IDRiD, donde el método presentado supera levemente al descrito en [35]. Cabe señalar que el de [35] utiliza una red neuronal preentrenada sobre pares co-registrados de angiografías por fluorescencia y FFO, obligando a contar con estos estudios para hacer el entrenamiento. Por el contrario, nuestro algoritmo parte de pesos aprendidos para clasificación de imágenes naturales (una tarea que difiere significativamente del objetivo del modelo final) y aún así es capaz de obtener resultados similares. En [35], además, se reportan los valores de AUC obtenidos por su propio modelo al entrenarlo a partir de los mismos pesos que el nuestro, observándose un rendimiento notoriamente inferior al alcanzado por nuestro algoritmo (AUC = 0.885 vs. 0.946, respectivamente). En el banco de datos MESSIDOR 2, nuestro método es sólo sobrepasado en términos de AUC y sensibilidad por el descrito en [5]. Sin embargo, dicho enfoque está entrenado sobre un grupo de 128.175 imágenes, mientras que el aquí presentado utiliza $\frac{1}{3}$ de esta cantidad. El método de Gargeya et al. [36], por otro lado, está entrenado sobre 75.137 imágenes (dos veces más que el nuestro), y obtiene un rendimiento inferior al aquí propuesto en términos de AUC y especificidad. Esto puede deberse tanto a que nuestro enfoque está entrenado sobre datos de múltiples orígenes como al efecto de la estrategia de aumentación de datos empleada, que permite diversificar la apariencia de las imágenes y fortalecer su capacidad de generalización. En el conjunto EyePACS, con más de 50.000 estudios para evaluación, nuestro algoritmo sobrepasa levemente el AUC reportado en [39], mientras que el de Quellec et al. [38] obtiene valores de AUC algo superiores al

nuestro. Este último, sin embargo, emplea un ensamble de 5 redes neuronales para obtener esos resultados, mientras que el nuestro utiliza una única red. Cuando el algoritmo de Quellec et al. [38] emplea sólo un modelo, su AUC cae a 0.944, que es inferior al reportado por nuestro algoritmo. Finalmente, en DR2 se observa que nuestro modelo obtiene un AUC superior al de Pires et al. [39], el cual recurre a entrenamientos múltiples a diferentes resoluciones de imagen para agregarle robustez a la respuesta de su red neuronal. Por el contrario, nuestro enfoque emplea un único entrenamiento sobre un volumen de datos diverso, combinado con una agresiva estrategia de aumentación de imágenes. En la comparación a nivel paciente realizada sobre MESSIDOR 2 y EyePACS (Tabla 4), se observa que el método de Pires et al. [39] funciona mejor en términos de AUC que el nuestro. No obstante, su enfoque para casos de múltiples imágenes por paciente se basa en utilizar la red neuronal como extractora de características para uno y otro ojo, y en emplear posteriormente un clasificador a nivel paciente basados en estos descriptores para brindar la respuesta final. Esto implica un proceso de aprendizaje adicional que en nuestro caso no se realiza, ya que nuestra respuesta a nivel paciente se obtiene simplemente tomando el máximo de las probabilidades obtenidas para cada imagen. En consecuencia, no es posible establecer a priori si estas diferencias de AUC se deben a que el método detallado en [39] utiliza una mejor red neuronal que la aquí presentada, o si aprovecha mejor sus respuestas gracias al clasificador a nivel paciente. En cualquier caso, resulta notorio cómo un modelo basado en una arquitectura de red neuronal convolucional estándar como es la ResNet puede igualar o superar resultados de alternativas más complejas basadas en un número significativamente mayor de parámetros gracias a un entrenamiento eficaz sobre un gran volumen de datos anotados.

Respecto al análisis cualitativo de las respuestas del algoritmo, se desprende de la Figura 8 que el modelo determina la ausencia de RD referible en fotografías centradas en el área macular mayormente analizando las arcadas vasculares

principales, el disco óptico y la mácula. En otro tipo de imágenes (por ejemplo, centradas en la papila o en regiones periféricas), se observa que el algoritmo brinda esta respuesta enfocándose sólo en el estado del disco óptico y las arcadas, ya que no tiene acceso a la zona macular. Los mapas también sugieren que las predicciones de RD referible se sustentan en una habilidad del modelo para detectar lesiones hemorrágicas y exudados blandos algodonosos y/o duros en la cercanía del área macular. Esto es destacable teniendo en cuenta que coincide con los mismos criterios aplicados en la práctica médica clínica, donde no solo son tenidos en cuenta como marcadores de diagnóstico sino también como factores predictores de amenaza de la función visual por su localización. Nótese además que estos criterios no fueron modelados explícitamente al momento de diseñar el algoritmo, sino que fueron aprendidos por éste a partir de los datos de entrenamiento. Por otro lado, los resultados en imágenes con otras patologías u observaciones oftalmológicas presentados en la sección 5.5, e ilustrados en la Figura 8 (c), permiten comprobar que en escenarios en los que los pacientes presentan condiciones diferentes a la RD, el algoritmo puede llegar a predecir RD referible basándose en la observación de estructuras o lesiones que son compatibles con la enfermedad o que se asemejan a ellas (por ejemplo, confundiendo un nevus con una hemorragia). Esto cobra relevancia en el contexto de un sistema telemédico, ya que permitiría tamizar casos con otras condiciones relevantes. No obstante, es necesario señalar que para la mayoría de las patologías urgentes del conjunto 1000Fundus el modelo no produce una respuesta alta uniforme y consistente, por lo que éstas deberían detectarse individualmente utilizando otros algoritmos específicos, que escapan al alcance de este primer trabajo.

Finalmente, un último punto a señalar tiene que ver con los resultados obtenidos sobre imágenes nacionales (sección 5.6). Aunque el número de estudios utilizados en esta evaluación es menor que el empleado en los demás análisis, los valores de AUC, sensibilidad y especificidad reportados son compatibles con los anteriores, lo

que indica su potencial aplicabilidad en un contexto clínico argentino. En este sentido es destacable la performance obtenida sobre los datos del Centro Oftalmológico Martínez, cuyas imágenes fueron adquiridas con un modelo de cámara no-midiátrica no utilizado para el entrenamiento del algoritmo.

6.2. Limitaciones del modelo y trabajos futuros

A pesar de la notoria precisión del algoritmo para el reconocimiento de la RD referible, existen algunas limitaciones que, de subsanarse, permitirían mejorar significativamente su rendimiento en contextos clínicos realistas.

Aunque los valores de AUC, sensibilidad y especificidad reportados en la sección 5.4 para reconocer casos más avanzados de RD son similares entre sí, resulta llamativo que el algoritmo sea más eficaz en reconocer la RDNP severa de la RDP. Esto puede deberse a que la cantidad de estudios proliferativos disponibles para entrenamiento es notoriamente inferior a la de los demás grados. Dado que el principal signo de este estadío es la presencia de neovasos, el algoritmo debe aprender por sí mismo a hallarlos, algo difícil de lograr si la cantidad de imágenes con presencia de neovascularizaciones es muy poca respecto al resto y sin supervisar al modelo indicándole la clase RDP de manera separada. En el ejemplo proliferativo de la Figura 8 (b) se observa que el algoritmo no concentra su atención sobre las neovascularizaciones en el disco óptico, sino sobre hemorragias. Este comportamiento se repite en un gran número de estudios proliferativos, lo que a priori puede ser indicativo de que, luego del entrenamiento, el modelo se inclina sobre todo a identificar este tipo de lesiones, junto a exudados y lesiones algodonosas, por sobre otras más específicas. Este inconveniente puede aliviarse modificando el modelo y reentrenándolo para predecir 3 clases (RD no referible, RDNP referible y RDP) en lugar de 2 (lo que lo obligaría a reconocer las diferencias entre RDNP referible y RDP durante el entrenamiento), o incorporando información explícita acerca de los vasos sanguíneos, introduciendo por ejemplo en la entrada la

segmentación vascular obtenida con algoritmos como el utilizado en [42]. No obstante, es importante destacar que la habilidad para reconocer casos de RDP no asegura per se que el tamizado cumpla con el objetivo principal de diagnóstico a tiempo, ya que es un estadio demasiado avanzado en el que es muy difícil lograr un tratamiento oportuno que garantice la conservación de la visión.

Vinculado con este último punto, es importante señalar que la retroalimentación que el modelo puede ofrecer al profesional mediante mapas de activación de clase no es tan precisa como la que ofrecen otros modelos orientados explícitamente en segmentar lesiones [34], que permiten delinearlas con precisión y luego utilizar las máscaras binarias resultantes para extraer estadísticas sobre las mismas (por ejemplo, su ubicación relativa respecto al disco óptico o la mácula, o el porcentaje de la imagen ocupado por ellas). Vale aclarar, sin embargo, que dichos enfoques suelen no ser robustos para el tamizado de casos de riesgo de RD. Según la comparativa realizada en las Tablas 3 y 4, el método de Zago et al. [34] obtiene resultados notoriamente peores para esta tarea respecto al aquí propuesto, en los conjuntos DDR, IDRID y MESSIDOR 2. Esto puede deberse a que dicho algoritmo identifica primero la probabilidad de existencia de lesiones rojas (microaneurismas y hemorragias) en parches de la imagen de entrada, y luego calcula como probabilidad de RD referible de la imagen al máximo de las probabilidades de lesión en todos los parches. En consecuencia, el diagnóstico está basado únicamente en la observación de un tipo específico de lesión, y los falsos positivos en la detección de las mismas reducen su especificidad notoriamente (ver Tabla 3, MESSIDOR 2). En un futuro se prevé agregarle mayor robustez a los mapas de activaciones del algoritmo propuesto modificándolo para que prediga no sólamente si la imagen corresponde a un caso de RD referible o no, sino también la presencia/ausencia de neovascularizaciones, hemorragias, exudados, etc. De esta forma, los mapas de activación derivados de cada respuesta de presencia/ausencia serán más localizados y permitirán retroalimentar al operador de la herramienta de una forma más clara.

Cuantificar de forma separada la incertidumbre del algoritmo ante casos ambiguos puede ser también de relevancia para brindar una mejor retroalimentación al profesional que eventualmente realizará el informe de los casos referibles. En la versión actual del algoritmo, las respuestas inciertas se corresponden con probabilidades cercanas al 50%, reflejando dudas en si optar por una u otra categoría. Brindar como salida adicional la entropía de ambas probabilidades (que toma valores altos cuando las probabilidades de referibilidad/no-referibilidad son similares) revelaría situaciones de incerteza de forma más clara [46], facilitando la interpretación médica posterior de la respuesta del algoritmo.

Finalmente, otra limitación a tratar explícitamente en un futuro tiene que ver con la aplicación del algoritmo sobre imágenes de mala calidad. En [41] se ha propuesto adaptar el modelo para predecir al mismo tiempo el grado de RD observada y la calidad del estudio. El resultado es un algoritmo capaz de mejorar sus resultados respecto a la versión original, ya que puede reconocer durante el entrenamiento cuándo una observación en la imagen se vincula realmente con la enfermedad en sí y cuándo se debe en realidad a un artefacto de captura. No obstante, diagnosticar la RD utilizando FFO de mala calidad es algo que no necesariamente deba modelarse desde la red neuronal en sí, ya que puede mitigarse más fácilmente recapturando los estudios que sean de calidad subóptima por problemas de captura. En este sentido, se han propuesto algoritmos para reconocer automáticamente estudios de mala calidad [19], que podrían aplicarse de forma adelantada para sugerir una recaptura antes de que el paciente abandone la sesión. En un futuro se prevé introducir algoritmos alternativos para este fin, que se ejecuten con anterioridad al de reconocimiento de RD referible para determinar de antemano si el estudio es o no procesable.

6.3. Implicancias clínicas del algoritmo para la teleoftalmología nacional

Aunque la teleoftalmología ha cobrado especial relevancia en términos internacionales en los últimos años, su aplicación en Argentina aún no ha alcanzado el nivel de madurez observado en otros lugares del mundo. Experiencias recientes como la red de teleoftalmología del Hospital El Cruce (HEC) creada en 2019 para cubrir el conurbano bonaerense [43] o la campaña con cámaras de fondo de ojo itinerantes en la provincia de La Pampa [44] han demostrado ser costo-efectivas para el reconocimiento temprano de la enfermedad, permitiendo alcanzar con diagnóstico a un mayor número de poblaciones.

En el caso del HEC, las imágenes son recolectadas por técnicos (mayormente enfermeros) desde hospitales provinciales y/o municipales y centros de atención primaria de 47 municipios distribuidos en 12 regiones sanitarias, utilizando 17 retinógrafos no-midriáticos instalados de manera permanente y otros itinerantes que visitan las localidades de menor densidad poblacional 2 veces al año. El Servicio de Oftalmología del hospital recibe los estudios y realiza un informe validado por 2 oftalmólogos en 48-72 horas, indicando si se trata de un caso patológico y cuál es la mejor estrategia de resolución, según guías consensuadas. Del total de población evaluada a noviembre de 2019, el 45.5% resultó patológico, y un 96% provino de nodos sin oftalmólogos [43]. La incorporación de un mayor número de nodos de captura implica de forma directa una mayor carga de trabajo para el centro de informes, que, tratándose del servicio encargado también del tratamiento, o bien verá limitada su capacidad de acción, o bien afectará la escalabilidad de la red. La experiencia de La Pampa es similar a la del HEC [44]. Este programa de teleoftalmología, iniciado en 2019, se basa en un retinógrafo itinerante que recorre zonas rurales de la provincia para capturar FFO. Las imágenes, obtenidas por enfermeros, se clasifican según su referibilidad de manera remota, y un oftalmólogo visita personalmente a quienes son señalados como referibles. Esta iniciativa ha demostrado lograr incrementar la tasa de adhesión de los pacientes diabéticos al

control anual oftalmológico de un 39.3% al 78.6% [44], en una población en la que se ha observado una prevalencia de la RD del orden del 21.5% [45].

En este tipo de redes, contar con un modelo como el aquí propuesto permitiría tamizar los estudios que requieren informe de manera automática. Escogiendo el punto operativo correcto en la curva ROC obtenida sobre los datos del HEC (Figura 9 (b)), es posible lograr una sensibilidad del 100% para una especificidad del 86.2%. Esto implica contar con un sistema capaz de reconocer a toda persona con RD referible, con una tasa de falsos positivos del 14%, que alivia entonces en un 86% la carga de trabajo del nodo de informes. Esto permite hacer un mejor uso del recurso humano disponible, a la vez que brindarle a los oftalmólogos retroalimentación mediante probabilidades de referibilidad y mapas de activación les facilita la generación del informe, guiándolos en el proceso y mejorando su eficiencia general.

7. Conclusiones

En este trabajo introdujimos un algoritmo de IA para el tamizado automático de casos de RD referible basado en redes neuronales convolucionales. Mediante el entrenamiento sobre múltiples bancos de datos públicos y gracias a una estrategia de preprocesamiento adecuada, fue posible obtener un modelo capaz de lograr tasas de detección superiores a las del estado del arte en numerosos conjuntos de evaluación. Según nuestro relevamiento bibliográfico, se trata del primer enfoque testeado a gran escala sobre más de 60.000 estudios y desarrollado íntegramente en nuestro país para el reconocimiento de la enfermedad. Esto ofrece una oportunidad relevante para Argentina en materia de tamizado poblacional masivo de la RD, que constituye la principal causa de ceguera preventible e irreversible en adultos en edad laboral en nuestro país. En tal sentido, se prevé en el mediano plazo integrar este modelo en una plataforma digital actualmente en desarrollo para asistir a la detección remota y eficiente de la RD.

Bibliografía

- [1] Silva JC, et al. Una evaluación comparativa de la ceguera y la deficiencia visual evitables en siete países latinoamericanos: prevalencia, cobertura y desigualdades. Rev Panam Salud Pública, 2015; 37(1): 21-28.
- [2] Lee DJ, et al. Dilated eye examination screening guideline compliance among patients with diabetes without a diabetic retinopathy diagnosis: the role of geographic access. BMJ Open Diabetes Res Care, 2014; 2(1): e000031.
- [3] Avidor D, Loewenstein A, Waisbord M, Nutman A. Cost-effectiveness of diabetic retinopathy screening programs using telemedicine: a systematic review. Cost Eff Resour Alloc, 2020; 18: 1-9.
- [4] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, (2015; 521(7553): 436-444.
- [5] Gulshan V, et al., Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 2016; 316(22): 2402-2410.
- [6] Orlando JI, et al., REFUGE challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Med Image Anal, 2020; 59: e101570.
- [7] Romo-Bucheli D, Erfurth US, Bogunović H. End-to-end deep learning model for predicting treatment requirements in neovascular AMD from longitudinal retinal OCT imaging. IEEE J Biomed, 2020; 24(12): 3456-3465.
- [8] Orlando JI, et al. Automated quantification of photoreceptor alteration in macular disease using optical coherence tomography and deep learning. Sci Rep, 2020; 10(1): 1-12.
- [9] Seeböck P. et al. Linking Function and Structure with ReSenseNet: Predicting Retinal Sensitivity from Optical Coherence Tomography using Deep Learning. Ophthalmology Retina, 2022; En prensa.

- [10] Abràmoff MD, Lavin, PT, Birch, M, Shah, N, Folk, JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ digital medicine*, 2018; 1(1): 1-8.
- [11] Xie Y, et al. Health economic and safety considerations for artificial intelligence applications in diabetic retinopathy screening. *Transl Vis Sci Technol*, 2020; 9(2): 22-22.
- [12] Teo ZL, et al. Do we have enough ophthalmologists to manage vision-threatening diabetic retinopathy? A global perspective. *Eye*, 2020; 34: 1255–1261.
- [13] Vouilloz, M, et al. Retinopatía diabética y agudeza visual en la primera consulta. 11º Jornadas Científicas del HEC, 2018. Online: <https://repositorio.hospitalelcruce.org/xmlui/handle/123456789/742?show=full>. Accedido por última vez el 27/3/2022.
- [14] Hong H, Mújica OJ, Anaya J, Lansingh VC, López E, Silva JC. The Challenge of Universal Eye Health in Latin America: distributive inequality of ophthalmologists in 14 countries. *BMJ Open*, 2016; 6:e012819.
- [15] Xie L, Yang S, Squirrell D, Vaghefi E. Towards implementation of AI in New Zealand national diabetic screening program: Cloud-based, robust, and bespoke. *PLoS one*, 2020; 15(4): e0225015.
- [16] Campbell, JP. Artificial intelligence to reduce ocular health disparities: Moving from concept to implementation. *Transl Vis Sci Technol*, 2021. 10(3): 19-19.
- [17] Li T, Gao Y, Wang K, Guo S, Liu H, Kang H. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Inf Sci*, 2019; 501: 511-522.
- [18] Galdran A, Dolz J, Chakor H, Lombaert H, Ben Ayed I. Cost-sensitive regularization for diabetic retinopathy grading from eye fundus images. *Lect Notes Comput Sci*, 2020; 12265: 665-674.
- [19] Fu H, et al. Evaluation of Retinal Image Quality Assessment Networks in Different Color-spaces. *Lect Notes Comput Sci*, 2019; 11764: 48-56.

- [20] Porwal P, et al. Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research. *Data*, 2018; 3(3): 25.
- [21] Castillo Benítez V, et al. Dataset from fundus images for the study of diabetic retinopathy. *Data Br*, 2021; 36: e107068.
- [22] Cen LP, et al. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nat Commun*, 2021. 12: 4828.
- [23] Pires R, et al. Beyond lesion-based diabetic retinopathy: a direct approach for referral. *IEEE J Biomed*, 2015; 21(1): 193-200.
- [24] Decencière E, et al. Feedback on a publicly distributed image database: the Messidor database. *Image Anal Stereol*, 2014; 33.3: 231-234.
- [25] Orlando JI, Prokofyeva E, del Fresno M, Blaschko MB. An ensemble deep learning based approach for red lesion detection in fundus images. *Comput Methods Programs Biomed*, 2018; 153: 115-127.
- [26] Kauppi T, et al. The DIARETDB1 diabetic retinopathy database and evaluation protocol. *BMVC*, 2007; 1: 1-10.
- [27] Otsu N. A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern: Syst*, 1979; 9(1): 62-66.
- [28] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proc. IEEE Int Conf Comput Vis*, 2016; 770-778.
- [29] Paszke A, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Adv Neural Inf Process Syst*, 2019; 32: 8024-8035.
- [30] Fu R, Hu Q, Dong X, Guo Y, Gao Y, Li B. Axiom-based grad-cam: Towards accurate visualization and explanation of CNNs. *arXiv preprint*, 2020; arXiv:2008.02312.
- [31] Fernandez, FG. TorchCAM: class activation explorer. Online: <https://github.com/frgfm/torch-cam>. Accedido por última vez el 27/3/2022.
- [32] Pedregosa, F. Scikit-learn: Machine learning in Python. *J Mach Learn Res*, 2011; 12: 2825-2830.

- [33] Wilson, EB. Probable inference, the law of succession, and statistical inference. *J Am Stat Assoc*, 1927; 22: 209-12
- [34] Zago GT, Andreão RV, Dorizzi B, Salles EOT. Diabetic retinopathy detection using red lesion localization and convolutional neural networks. *Comput Biol Med*, 2020; 116: e103537.
- [35] Hervella A, Rouco J, Novo J, Ortega M. Multimodal image encoding pre-training for diabetic retinopathy grading. *Comput Biol Med*, 2022; 143: e105302.
- [36] Gargya R, Leng T. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*, 2017; 124(7): 962-969.
- [37] Voets M, Møllersen K, Bongo LA. Reproduction study using public data of: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *PloS one*, 2019; 14(6): e0217541.
- [38] Quellec G, Charrière K, Boudi Y, Cochener B, Lamard M. Deep image mining for diabetic retinopathy screening. *Med Image Anal*, 2017; 39: 178-193.
- [39] Pires R, Avila S, Wainer J, Valle E, Abràmoff, MD, Rocha A. A data-driven approach to referable diabetic retinopathy detection. *Artif Intell Med*, 2019; 96: 93-106.
- [40] Abràmoff MD. et al. Foundational Considerations for Artificial Intelligence Using Ophthalmic Images. *Ophthalmol*, 2022; 129(2): 14-32.
- [41] Zhou K, Gu Z, Li A, Cheng J, Gao S, Liu J. Fundus image quality-guided diabetic retinopathy grading. *Lect Notes Comput Sci*, 2018; 11039: 245-252.
- [42] Orlando JI, van Keer K, Barbosa Breda J, Manterola HL, Blaschko MB, Clausse A. Proliferative diabetic retinopathy characterization based on fractal features: Evaluation on a publicly available dataset. *Med Phys*, 2017; 44(12): 6425-6434.
- [43] Koch, et al. Estrategias innovadoras para mejorar los cuidados a personas con Enfermedades Crónicas. Reporte técnico, 2019. Ministerio de Salud de la Provincia de Buenos Aires.

- [44] Ortiz-Basso T, Gomez PV, Boffelli A, Paladini A. Programa de teleoftalmología para prevención de la ceguera por diabetes en una zona rural de la Argentina. Revista de la Facultad de Ciencias Médicas de Córdoba, 2022; 79(1): 10-14.
- [45] Ortiz-Basso T, Boietti BR, Gómez PV, Boffelli AD, Paladini AA. Prevalence of diabetic retinopathy in a rural area of Argentina. Medicina, 2022; 82(1): 99-103.
- [46] Araújo T, et al. DR|GRADUATE: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. Med Image Anal, 2020; 63: e101715.

Tabla 1. Construcción de los conjuntos de entrenamiento y validación y de los múltiples conjuntos de test a partir de bases de datos públicas y de dos conjuntos nacionales (Martínez y HEC), indicando la cantidad de muestra sin RD referible y con RD referible y el total de muestras (N) de cada uno.

Nombre	Muestras			Entrenam.	Validación	Test
	Sin RD referible	Con RD referible	N			
DDR	6.896	5.626	12.522	6.260	2503	3.759
DeepDRID	900	700	1.600	1.200	0	400
APTOS2019	2.175	1.487	3.662	3.662	0	0
EyePACS	71.548	17.154	88.702	28.098	7026	53.576
IDRiD	193	323	516	372	40	103
FCM-UNA	191	566	757	0	0	757
1000Fundus	56	88	144	0	0	144
DR2	337	98	435	0	0	435
MESSIDOR 2	1.287	457	1.744	0	0	1.744
DIARETDB1	43	46	89	0	0	89
Martínez	454	30	484	0	0	484
HEC	26	9	35	0	0	35
Total	89.433	27.735	117.168	39.592	9.569	61.526

Tabla 2. Operaciones de aumentación de imágenes aplicadas para el entrenamiento del algoritmo y parámetros empleados para cada una. Para cada transformación, se referencia a la figura de ejemplo demostrando el resultado de su aplicación.

Operación	Descripción	Probabilidad de aplicación	Parámetros	Ejemplo
Fluctuación de color	Cambios aleatorios en: <ul style="list-style-type: none"> • Brillo (b) • Contraste (c) • Saturación (s) 	0.25	$b \in [-25\%, +25\%]$ $c \in [-25\%, +25\%]$ $s \in [-25\%, +25\%]$	Figura 4 (b)
Espejado horizontal	Volteo aleatorio de la imagen de izquierda a derecha.	0.7		Figura 4 (c)
Espejado vertical	Volteo aleatorio de la imagen de arriba abajo.	0.5		Figura 4 (d)
Rotación	Rotación en un ángulo aleatorio.	1.0	Ángulo $\in [-180^\circ, +180^\circ]$	Figura 4 (e)
Escalado	Cambio en el tamaño de la imagen seguido de un recorte (si agranda) o de completado con nulos (si encoge).	1.0	Escala $\in [-40\%, +10\%]$	Figura 4 (f)

Tabla 3. Resultados obtenidos por el modelo propuesto y otros de la literatura en los conjuntos de test, evaluados en términos de área bajo curva ROC (AUC), sensibilidad (Se) y especificidad (Esp). Se incluyen los intervalos de confianza (IC) del 95% para el modelo propuesto y para aquellos que los reportan en sus respectivos artículos.

Conjunto	Método	AUC (IC)	Se (IC)	Esp (IC)
DDR	Zago et al. 2020 [34]	0.833 (0.819 - 0.846)	-	-
	<i>Nuestro modelo</i>	0.965 (0.960 - 0.970)	0.749 (0.728 - 0.769)	0.978 (0.971 - 0.984)
IDRiD	Zago et al. 2020 [34]	0.796 (0.715 - 0.892)	-	-
	Hervella et al. 2022 [35]	0.944	-	-
	<i>Nuestro modelo</i>	0.949 (0.914 - 0.980)	0.828 (0.718 - 0.901)	0.897 (0.764 - 0.959)
MESSIDOR 2	Zago et al. 2020 [34]	0.944 (0.925 - 0.966)	0.900 (0.860 - 0.961)	0.8700 (0.863 - 0.871)
	Gulshan et al. 2016 [5]	0.990 (0.986 - 0.995)	0.961 (0.924 - 0.983)	0.939 (0.924 - 0.953)
	Gargeya et al. 2017 [36]	0.940	0.930	0.870
	Voets et al. 2019 [37]	0.853 (0.835 - 0.871)	0.818	0.687
	<i>Nuestro modelo</i>	0.973 (0.967 - 0.979)	0.895 (0.863 - 0.920)	0.941 (0.927 - 0.953)
EyePACS	Quellec et al. 2017 [38]	0.954	-	-
	Pires et al. 2019 [39]	0.946	-	-
	<i>Nuestro modelo</i>	0.951 (0.949 - 0.954)	0.732 (0.723 - 0.740)	0.979 (0.978 - 0.980)
DR2	Pires et al. 2019 [39]	0.963 (0.938 - 0.981)	-	-
	<i>Nuestro modelo</i>	0.974 (0.962 - 0.985)	0.847 (0.763 - 0.905)	0.961 (0.935 - 0.977)
1000Fundus	<i>Nuestro modelo</i>	1.000	1.000	0.964

		(1.000 - 1.000)	(0.958 - 1.000)	(0.879 - 0.990)
DeepDRID	<i>Nuestro modelo</i>	0.959 (0.944 - 0.972)	0.883 (0.828 - 0.922)	0.868 (0.817 - 0.907)
DIARETDB1	<i>Nuestro modelo</i>	0.981 (0.956 - 0.999)	0.957 (0.855 - 0.988)	0.930 (0.814 - 0.976)
FCM-UNA	<i>Nuestro modelo</i>	0.986 (0.980 - 0.992)	0.882 (0.852 - 0.906)	0.990 (0.963 - 0.997)
Martínez	<i>Nuestro modelo</i>	0.955 (0.927 - 0.980)	0.800 (0.627 - 0.905)	0.934 (0.907 - 0.953)
HEC	<i>Nuestro modelo</i>	0.961 (0.900 - 1.000)	1.000 (0.610 - 1.000)	0.862 (0.694 - 0.945)

Tabla 4. Resultados obtenidos en los datos de test de EyePACS, el conjunto MESSIDOR 2 y los datos de validación de DeepDRID, evaluando a nivel de paciente en términos de área bajo curva ROC (AUC), sensibilidad (Se) y especificidad (Esp). Se incluyen los intervalos de confianza (IC) del 95% para el modelo propuesto y aquellos que los reportan en sus respectivos artículos.

Conjunto	Método	AUC (IC)	Se (IC)	Esp (IC)
MESSIDOR 2 (N = 870 pacientes)	Pires et al. 2019 [39]	0.982 (0.974 - 0.989)	-	-
	Zago et al. 2020 [34]	0.944 (0.927 - 0.965)	-	-
	Nuestro modelo	0.970 (0.960 - 0.979)	0.907 (0.866 - 0.936)	0.920 (0.896 - 0.939)
EyePACS (N = 26.788 pacientes)	Pires et al. 2019 [39]	0.955 (0.951 - 0.958)	-	-
	Zago et al. 2020 [34]	0.821 (0.812 - 0.829)	-	-
	Nuestro modelo	0.948 (0.945 - 0.951)	0.773 (0.762 - 0.783)	0.969 (0.966 - 0.971)
DeepDRID (N = 100 pacientes)	Nuestro modelo	0.980 (0.957 - 0.999)	0.980 (0.895 - 0.996)	0.800 (0.670 - 0.888)

Tabla 5. Resultados obtenidos en múltiples conjuntos evaluando la capacidad del modelo de distinguir a personas sin RD referible respecto a grados más avanzados de la enfermedad, en términos de área bajo curva ROC (AUC), sensibilidad (Se) y especificidad (Esp). Se incluyen los intervalos de confianza (IC) del 95% para el modelo propuesto.

Conjunto	Sin RD referible vs.	AUC (CI)	Se (CI)	Esp (CI)
DDR	RDNP moderada	0.959 (0.953 - 0.964)	0.714 (0.689 - 0.737)	0.978 (0.971 - 0.984)
	RDNP severa	1.000 (0.999 - 1.000)	1.000 (0.949 - 1.000)	
	RDP	0.990 (0.986 - 0.993)	0.858 (0.812 - 0.894)	
IDRiD	RDNP moderada	0.927 (0.872 - 0.971)	0.750 (0.579 - 0.867)	0.897 (0.764 - 0.959)
	RDNP severa	0.956 (0.886 - 1.000)	0.895 (0.686 - 0.971)	
	RDP	0.990 (0.970 - 1.000)	0.923 (0.667 - 0.986)	
MESSIDOR 2	RDNP moderada	0.966 (0.958 - 0.974)	0.865 (0.825 - 0.897)	0.941 (0.927 - 0.953)
	RDNP severa	1.000 (1.000 - 1.000)	1.000 (0.951 - 1.000)	
	RDP	0.986 (0.970 - 0.997)	0.971 (0.855 - 0.995)	
EyePACS	RDNP moderada	0.940 (0.937 - 0.943)	0.672 (0.662 - 0.682)	0.979 (0.978 - 0.980)
	RDNP severa	0.991 (0.989 - 0.994)	0.951 (0.937 - 0.961)	
	RDP	0.988 (0.985 - 0.990)	0.900 (0.882 - 0.916)	
DeepDRID	RDNP moderada	0.941 (0.919 - 0.962)	0.837 (0.748 - 0.899)	0.868 (0.817 - 0.907)
	RDNP severa	0.979 (0.967 - 0.989)	0.926 (0.839 - 0.968)	
	RDP	0.971 (0.950 - 0.988)	0.950 (0.764 - 0.991)	

FCM-UNA	RDNP moderada	0.927 (0.893 - 0.958)	0.700 (0.592 - 0.789)	0.990 (0.963 - 0.997)
	RDNP severa	0.996 (0.993 - 0.998)	0.869 (0.812 - 0.911)	
	RDNP muy severa	0.999 (0.998 - 1.000)	0.981 (0.935 - 0.995)	
	RDP	0.995 (0.990 - 0.999)	0.920 (0.845 - 0.961)	
	RDP avanzada	0.994 (0.988 - 1.000)	0.904 (0.835 - 0.945)	

Figura 1. Algunos ejemplos de imágenes sin RD referible (marco verde) y con RD referible (marco violeta) extraídas de las bases de datos públicas utilizadas para construir los datos de entrenamiento, validación y test del algoritmo propuesto.

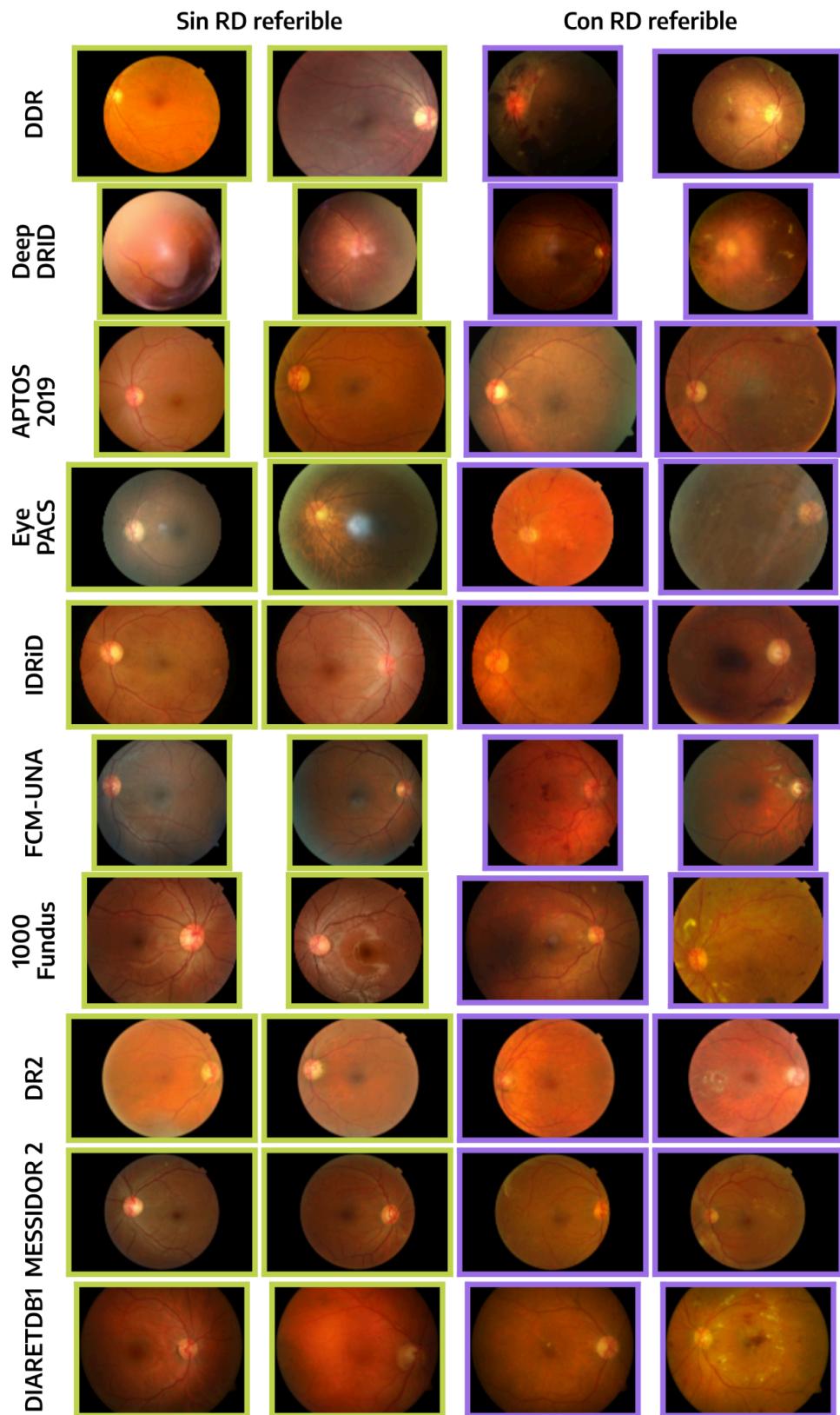
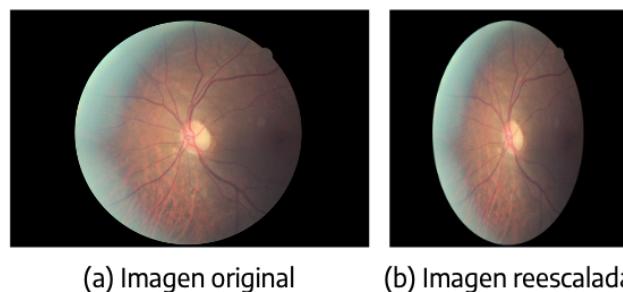


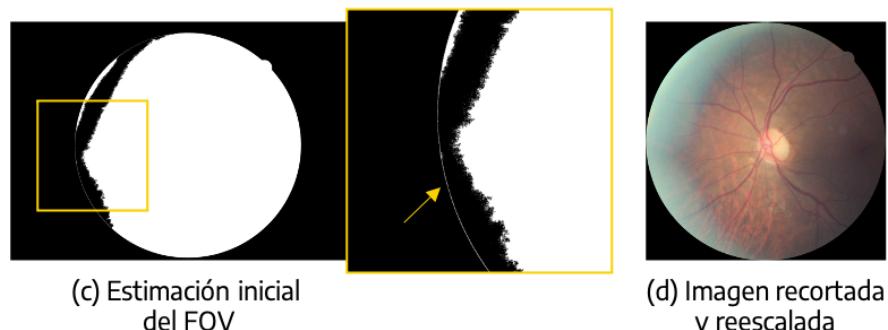
Figura 2. Arriba: ejemplo de una imagen preparada para entrenamiento sin preprocesamiento. Abajo: estrategia de preprocesamiento propuesta: 1. Recorte en el área del FOV y reescalado a 512 x 512 píxeles. 2. Mejora del contraste basada en la corrección de la estimación del FOV, la estimación de un fondo y su sustracción.

Sin preprocesamiento



Preprocesamiento

1. Recorte y reescalado



2. Mejora de contraste

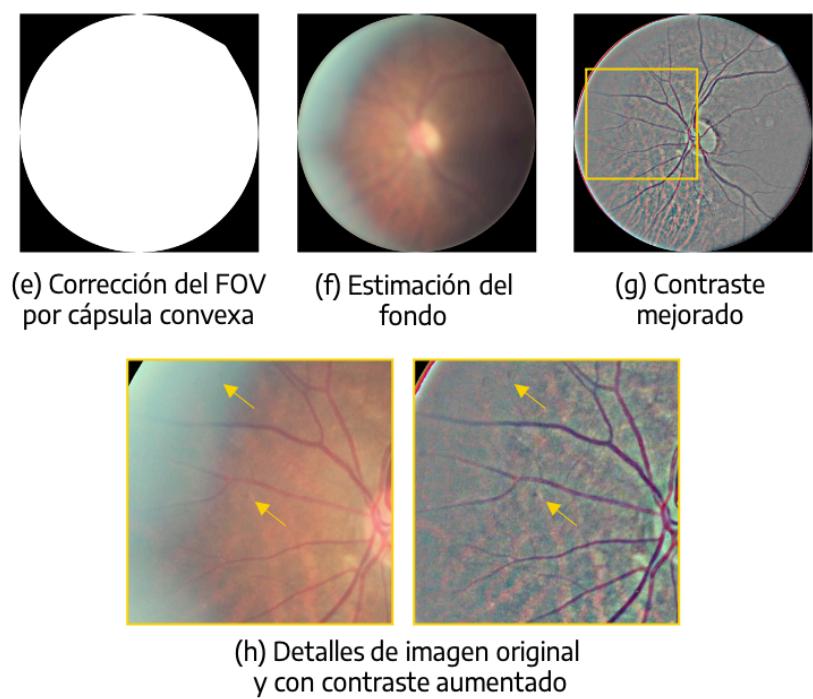


Figura 3. Representación esquemática de la arquitectura de red neuronal utilizada para resolver el problema de detección automática de retinopatía diabética referible. Las flechas corresponden a conexiones por adelantamiento, donde las representadas con líneas punteadas incluyen además una operación de complementado con ceros.

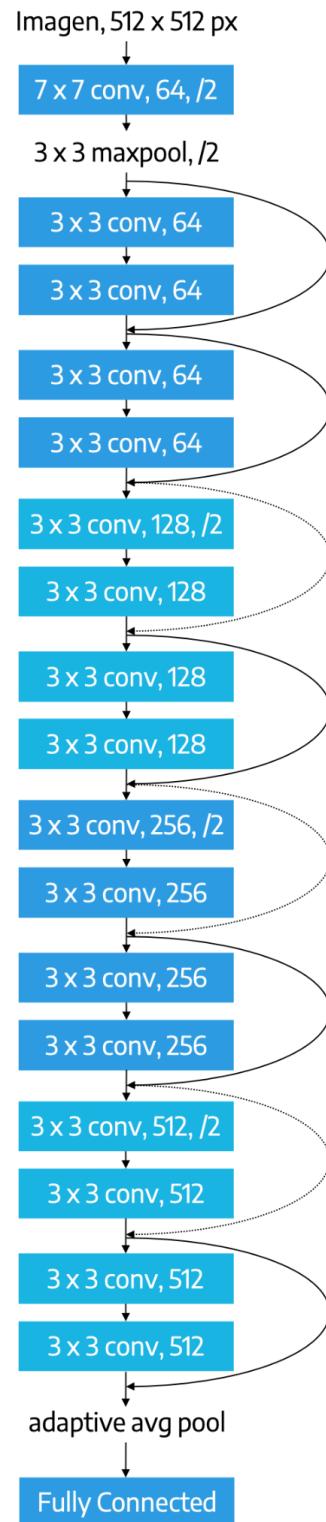


Figura 4. Ejemplo de las transformaciones sucesivas aleatorias aplicadas sobre las imágenes de entrenamiento para la aumentación de datos.

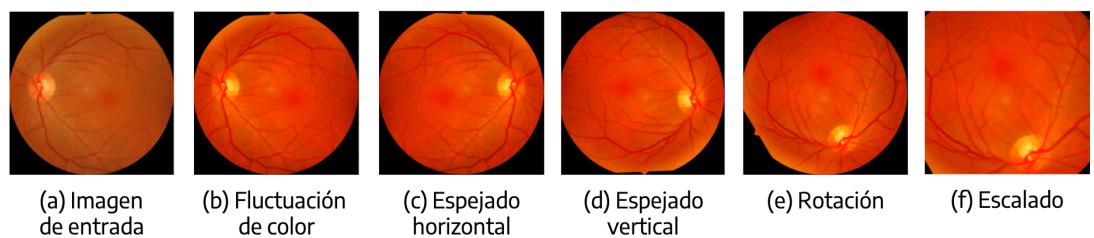


Figura 5. Curvas ROC obtenidas sobre imágenes de diferente calidad en dos subconjuntos de los datos de test de EyePACS (a y b) y en el conjunto de validación de DeepDRID (c).

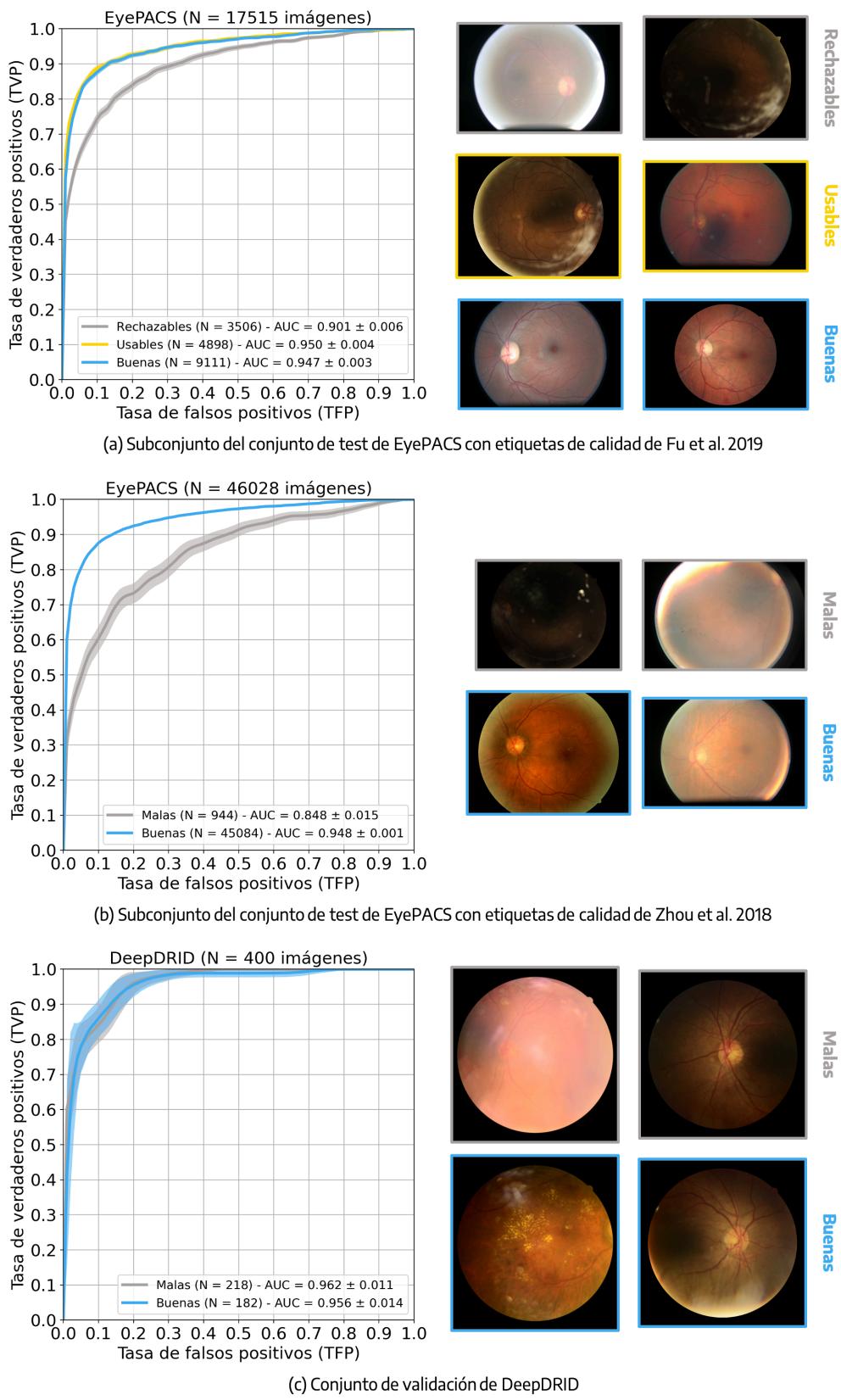


Figura 6. Curvas ROC obtenidas al diferenciar personas sin RD referible respecto de pacientes con niveles crecientes de RD no proliferativa (RDNP) y proliferativa (RDP). Los valores de N en las leyendas indican la cantidad de imágenes del grado correspondiente incluidas en el testeo, vs. aquellas sin RD referible.

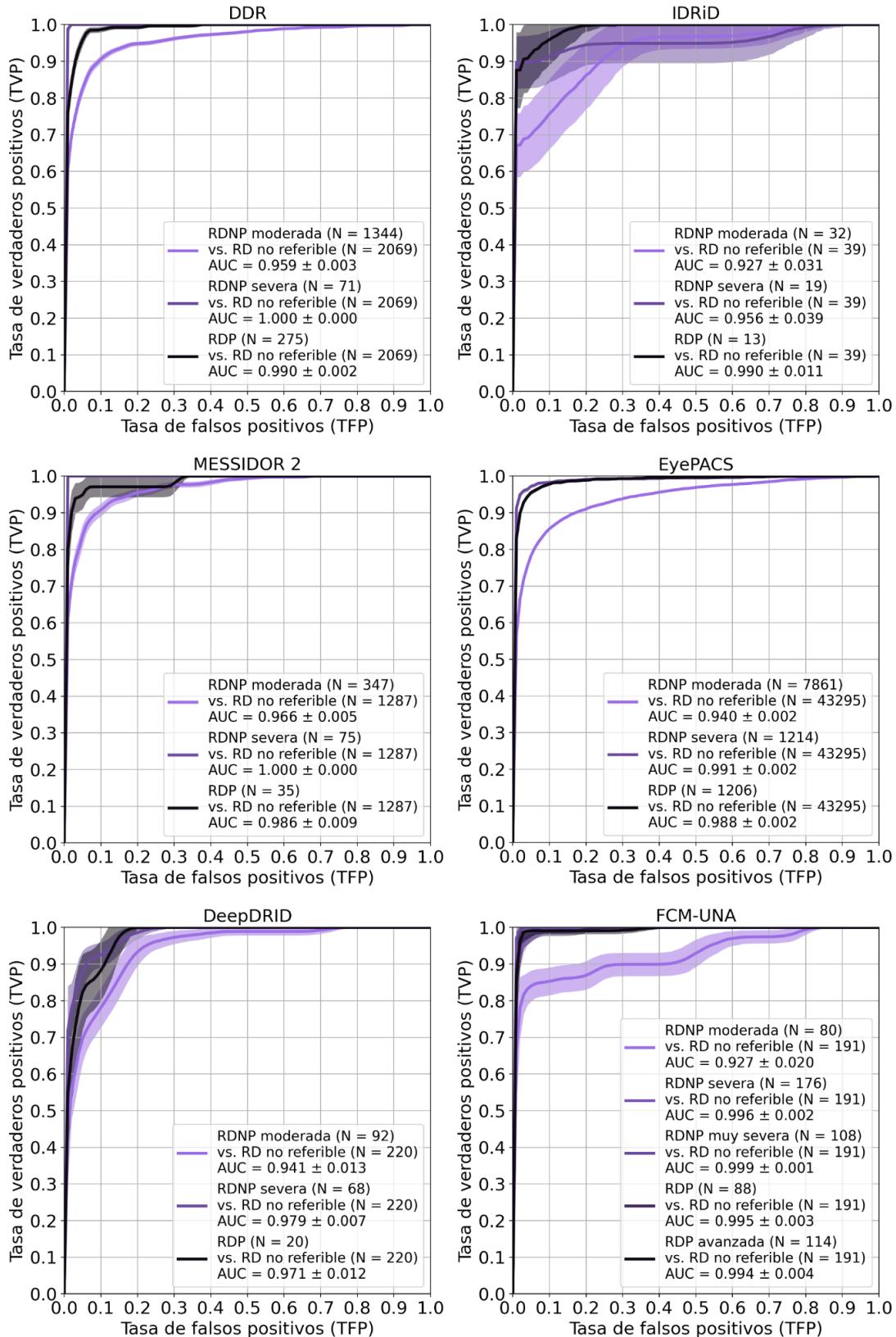


Figura 7. Distribución de las probabilidades de RD referible predichas por el algoritmo sobre imágenes del conjunto 1000Fundus con diversas observaciones oftalmológicas.

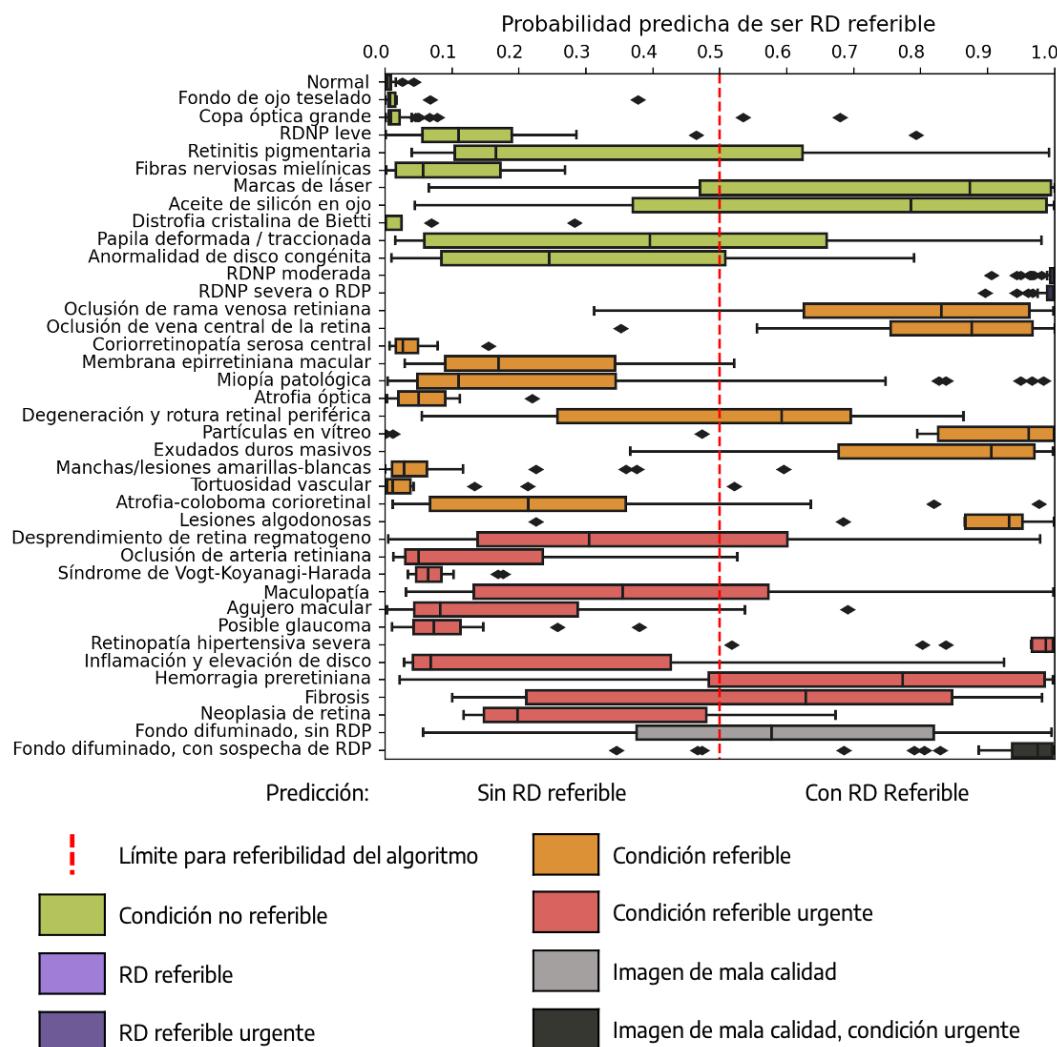


Figura 8. Algunos ejemplos de resultados cualitativos generados por la red neuronal propuesta, sobre diversos escenarios de evaluación. (a) 4 imágenes de un mismo paciente, 2 por cada ojo, todas señaladas como de mala calidad en DeepDRID. (b) Imágenes con diferentes grados de RD observada, según el conjunto FCM-UNA. (c) Imágenes de 1000Fundus con diversas observaciones.

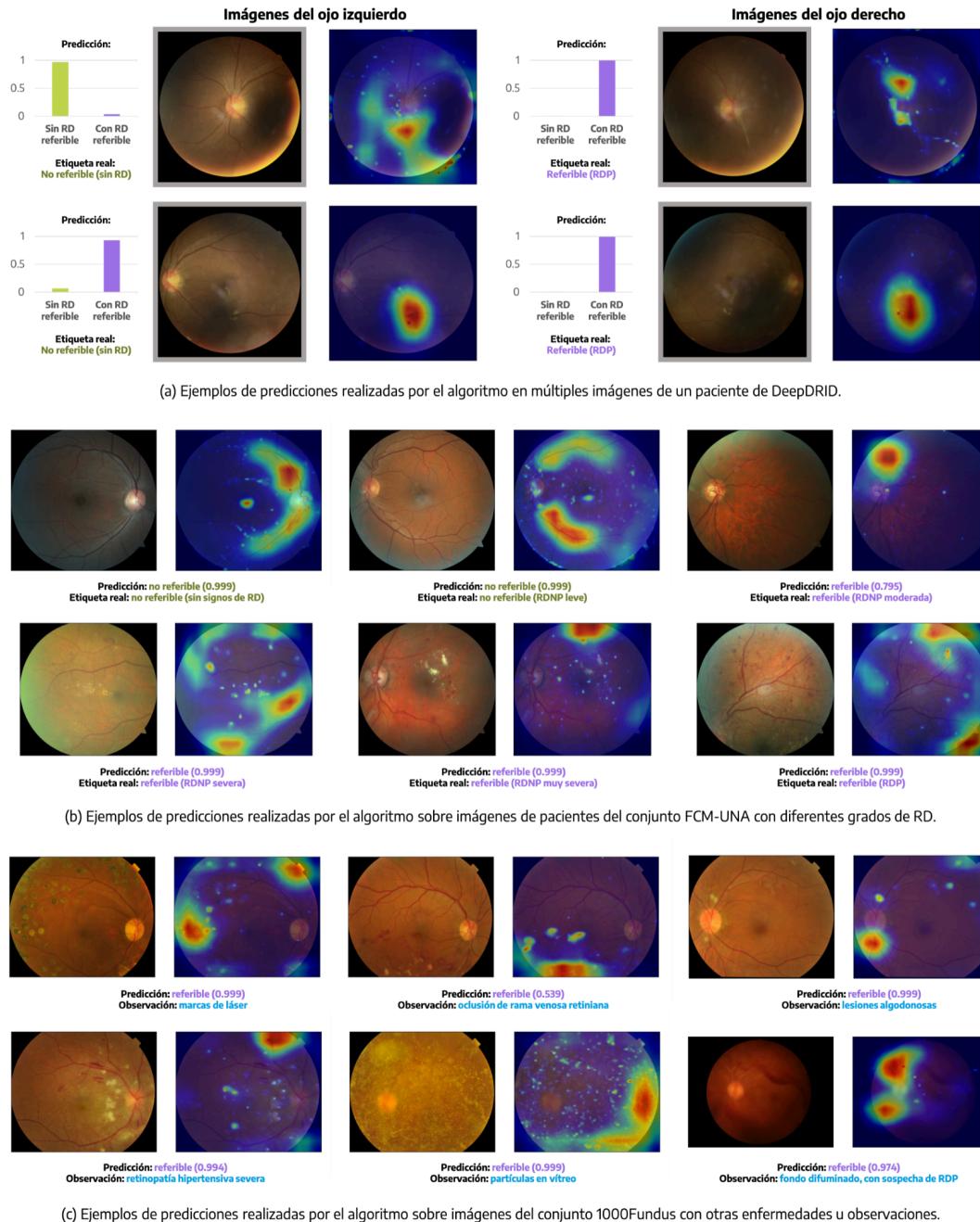
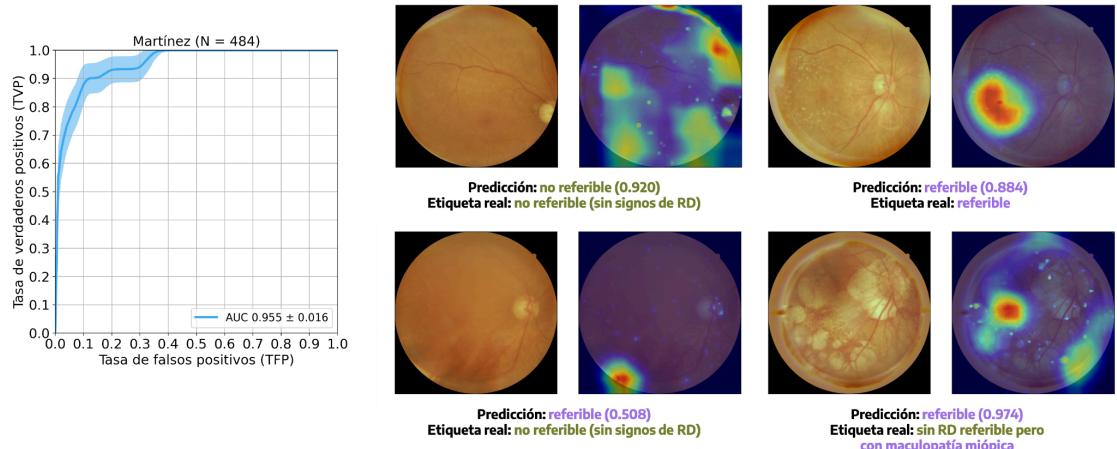
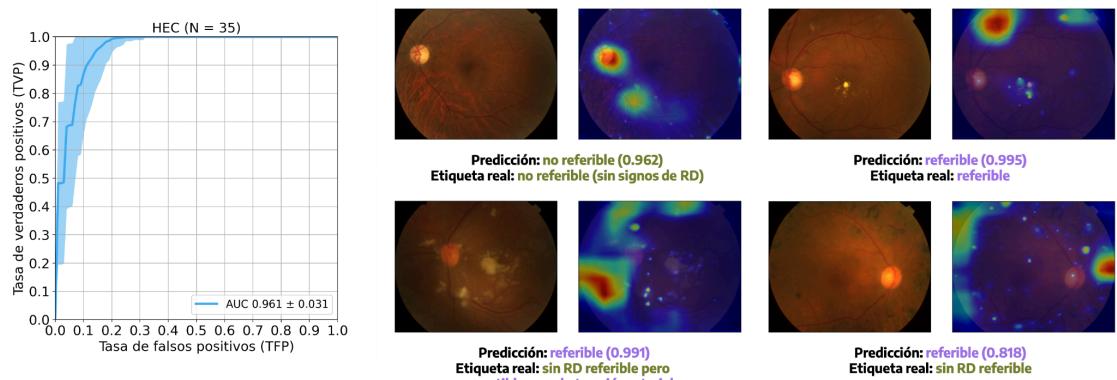


Figura 9. Resultados cuantitativos y cualitativos obtenidos sobre los conjuntos de datos nacionales (a) Martínez y (b) HEC. Izquierda: curvas ROC y valores de área bajo curva (AUC). Derecha: ejemplos de imágenes, predicciones, etiquetas reales y sus correspondientes mapas de activación asociados.



(a) Resultados sobre imágenes del conjunto Martínez.



(b) Resultados sobre imágenes del conjunto HEC.