Evaluación retrospectiva de un primer modelo de inteligencia artificial argentino para tamizaje automático de retinopatía diabética referible a partir de fotografías de fondo de ojo

José Ignacio Orlando, Tomás Castilla, Alejandro Koch, Ignacio Larrabide, Marcela Martínez y Mercedes Leguía.

Resumen:

Objetivos: Evaluar la efectividad de una red neuronal convolucional para la identificación de casos de retinopatía diabética (RD) referible a partir de fotografías de fondo de ojo, entrenada con 39.592 estudios recolectados de conjuntos públicos.
Materiales y Métodos: Se realizó un estudio observacional retrospectivo sobre 61.525 retinografías no empleadas para entrenamiento. 61.007 imágenes son internacionales y de origen público, utilizadas comúnmente para evaluar estos algoritmos, y están agrupadas en 9 conjuntos diferentes. Las restantes 519 fueron recolectadas retrospectivamente de las bases de datos clínicas del Centro de Oftalmología Martínez (Pehuajó, Argentina) y del Hospital de Alta Complejidad El Cruce (Florencio Varela, Argentina). Para cada imagen, se cuenta con una etiqueta asociada indicando si el caso corresponde a un paciente con signos de RD referible o no. En el caso de los datos de Argentina, estas etiquetas fueron asignadas manualmente por dos oftalmólogas expertas. Las rpobabilidades de salida del algoritmo se compararon con las etiquetas manuales utilizando curvas ROC y cuantificando tanto área bajo la curva (AUC) como valores de sensibilidad (SE) y especificidad (ESP) y sus intervalos de confianza (IC 95%). Para SE y ESP se utilizaron tres puntos operativos diferentes, seleccionados utilizando 9.569 estudios que no se emplearon ni para entrenar ni para evaluar el algoritmo. Estos puntos corresponden a umbrales de probabilidad que aseguran SE alta (PO1), ESP alta (PO2) o que consideran referibles a los casos con un 50% de probabilidad asignada (PO50%). Para analizar cualitativamente la respuesta del modelo, se estudiaron manualmente las regiones que el algoritmo tuvo en cuenta utilizando mapas de calor obtenidos mediante la técnica XGrad-Cam.
Resultados: El modelo reportó un AUC = 0.954 (0.952-0.956). Para el PO1, se obtuvieron SE = 95.5% (95.1%-95.8%) y ESP = 70.2% (69.8%-70.6%), para PO2 SE = 74.8% (74.0%-75.5%) y ESP = 0.97.8% (97.6%-97.9%) y para PO50% SE = 75.2% (74.4%-75.9%) y ESP = 97.7% (97.5%-97.8%). En los casos correctamente clasificados como referibles, los mapas de calor indicaron que el algoritmo analizó mayormente la presencia de hemorragias y/o exudados duros y algodonosos, mientras que en los correctamente clasificados como no referibles las activaciones más notorias se reconocen en las regiones de la mácula, el nervio óptico y las arcadas vasculares. Los casos erróneos, por otro lado, se asocian mayormente a problemas de captura tales como suciedad en la lente y artefactos, o a la presencia de otras enfermedades con signos similares a los de la RD.
Conclusiones: El modelo demostró valores altos de SE y ESP y respuestas cualitativas compatibles con las observaciones normalmente realizadas por los fotalmólogos. Se espera continuar recolectando datos para realizar una evaluación más exhaustivas sobre datos nacionales, con el propósito de integrar este algoritmo a una plataforma digital argentina actualmente en desarrollo para el tamizaje de la RD.

Resumen extendido:

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 ese control [2], en muchos casos recurriendo a la consulta ya en estado de ceguera legal, sin posibilidad de ser tratados eficazmente. Las redes de telemedicina oftalmológica ofrecen una solución a este inconveniente utilizando fotografías de fondo de ojo (FFO), por su costo-efectividad y facilidad de captura. Sin embargo, estos enfoques tienen una capacidad de escalabilidad limitada, ya que al incrementarse la afluencia de personas diabéticas a los nodos de captura de estudios o al incorporarse múltiples nodos nuevos puede saturarse el 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) han revolucionado el campo de la telemedicina oftalmológica. Estos enfoques son capaces por ejemplo de detectar y luego filtrar automáticamente los casos que requieren ser analizados por un profesional, reduciendo significativamente la carga de trabajo de los centros de informes [3]. Sin embargo, las soluciones comerciales existentes, mayormente desarrolladas en Estados Unidos, Asia y Europa, suelen requerir de cámaras específicas o están entrenados para etnias y poblaciones determinadas, por lo que fallan al aplicarse sobre estudios de otros orígenes [4]. Por otro lado, no existen en Argentina desarrollos nacionales en esta dirección, lo que hoy por hoy detiene la posibilidad de planificar estrategias de tamizaje 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 la rigidez de estas mismas herramientas.

Para suplir esta necesidad concreta, los autores de este trabajo desarrollaron un primer prototipo nacional de IA basado en redes neuronales convolucionales para el reconocimiento automático de la RD a partir de FFO. 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). 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. En este trabajo se presentan los resultados de un estudio de evaluación retrospectivo y observacional realizado con el propósito de estudiar su efectividad al aplicarse sobre un gran volumen de datos clínicos.

2. Materiales y métodos

2.1. Materiales

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 cuentan con etiquetas indicando si cada imagen presenta o no RD referible, o que proveen 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, se asociaron los casos sin RD o con RD no proliferativa (RDNP) leve a la clase “no referible”, mientras que las RDNP moderadas y severas o los casos de RD proliferativa (RDP) se asociaron a la clase “referible”. En total se recolectaron 117.168 estudios, capturados de forma midriática y no-midriática con más de 41 tipos diferentes de cámaras retinales (incluyendo modelos Topcon D7000, TRC NW6, NW8, 50DX y NW400, Zeiss FF450 Plus y Visucam 500 y Kowa VC-10alpha, entre otros) y con ángulos de apertura entre 35º y 50º. Del total de imágenes, 27.735 corresponden a casos con RD referible y 89.433 a casos no referibles.

Las imágenes fueron separadas en conjuntos disjuntos de entrenamiento, validación y test. El conjunto de entrenamiento fue utilizado para entrenar el modelo de IA, y consistió en 39.592, recogidos de las particiones predeterminadas de entrenamiento en los bancos públicos DDR (6.260), DeepDRID (1.200), APTOS2019 (3.662) y EyePACS (28.098). El de validación fue utilizado para calibrar los parámetros del algoritmo y ajustar los puntos operativos, y fue construido con 9.569 imágenes recolectadas del conjunto de validación predeterminado de DDR (2.503) y extraídas aleatoriamente de los predeterminados como de entrenamiento en EyePACS (7.026) e IDRiD (40). Finalmente, el conjunto de test fue utilizado para evaluar la performance del algoritmo, contando con un total de 61.526 imágenes provenientes de los conjuntos de test predeterminados de DDR (3.759), EyePACS (53.576), IDRiD (103), las predeterminadas como de validación en DeepDRID (400), con todas las de los conjuntos FCM-UNA (757), DR2 (435), MESSIDOR 2 (1.748) y DIARETDB1 (89), y con las señaladas en 1000Fundus como normales o con algún grado de RD (144). Además, se incorporaron a los datos de test 519 imágenes recolectadas retrospectivamente de las bases de datos del Centro Oftalmológico Martínez de Pehuajó (conjunto Martínez) y del 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, para la evaluación del algoritmo sobre estudios argentinos. El protocolo para utilizar estos estudios en un análisis post-hoc fue avalado oportunamente por el Comité de Ética del Hospital El Cruce (dictamen 25/2020), y adhiere a los términos de la Declaración de Helsinki. Las imágenes fueron debidamente anonimizadas en cada clínica antes de ser transferidas al equipo científico encargado de su evaluación, para resguardar la identidad de los pacientes. El conjunto Martínez está constituido por 484 imágenes obtenidas con un

dispositivo Cristal Vue NFC-700, no midriático y con 45º de FOV, 30 de las cuales presentan signos de referibilidad de RD. 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, que incluye 9 imágenes de personas con RD referible. Las capturas 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 para RD referible o no como para otras observaciones diagnósticas.

2.2. Métodos

El algoritmo de IA fue evaluado cuantitativamente utilizando curvas ROC y el área bajo cada curva (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 a partir de predicciones binarias obtenidas en tres puntos operativos diferentes, seleccionados utilizando 9.569 estudios que no se emplearon ni para entrenar ni para evaluar el algoritmo. Estos puntos corresponden a umbrales de probabilidad que aseguran SE alta (PO1, probabilidad > 2.18%), ESP alta (PO2, probabilidad > 51.23%) o que consideran referibles a los casos con un 50% de probabilidad asignada (PO50%, probabilidad > 50%). 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 Python[1] que aplica el método de Wilson [6]. Para el estudio cualitativo de los resultados del modelo se extrajeron mapas de activación de clases mediante la técnica XGrad-Cam [7], empleando la implementación de la librería Torchcam [8]. 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. Para su representación gráfica se utilizó un mapa de calor, donde las regiones con tonalidades rojas se asocian a activaciones altas (regiones con un impacto positivo en la respuesta) y las azules a activaciones bajas (regiones de impacto negativo).

3. Resultados

3.1. Resultados cuantitativos

El modelo reportó sobre el conjunto completo de datos de test un AUC = 0.954 (0.952-0.956). En el PO1, se obtuvo una SE = 95.5% (95.1%-95.8%) para una ESP = 70.2% (69.8%-70.6%), mientras que en el PO2 el algoritmo reportó una SE = 74.8% (74.0%-75.5%) para una ESP = 97.8% (97.6%-97.9%) y en el PO50% una SE = 75.2% (74.4%-75.9%) para una ESP = 97.7% (97.5%-97.8%). La Tabla 1 presenta los resultados separados para cada subconjunto de datos particular, utilizando el PO50% para calcular los valores de SE y SP.

3.2. Resultados cualitativos

En la Figura 1 (a) se incluyen resultados cualitativos obtenidos por el modelo para diferentes grados de RD sobre imágenes del conjunto FCN-UNA bien clasificadas. Se observa que en los casos correctamente clasificados 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, no se observan activaciones en la zona con neovascularizaciones, sino sobre algunas lesiones rojas.

En la Figura 2 (b) se ejemplifican 4 casos del conjunto Martínez (parte superior) y 2 del conjunto HEC. El primero de los casos del conjunto Martínez corresponde a un paciente sin RD 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

[1] https://gist.github.com/maidens/29939b3383a5e57935491303cf0d8e0b

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 anormalidades presentes en la región foveal. El primero de los ejemplos del HEC 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. El caso de la derecha, por otro lado, 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 los exudados como signos de la enfermedad.

3.2 Discusión

Aunque la teleoftalmología ha cobrado especial relevancia en términos internacionales en los últimos años, su aplicación en Argentina no ha alcanzado la madurez observada 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 [10] o la campaña con cámaras de fondo de ojo itinerantes en la provincia de La Pampa [11] han demostrado ser costo-efectivas para el reconocimiento temprano de la enfermedad, permitiendo alcanzar con diagnóstico a un mayor número de poblaciones.

En este tipo de enfoques, contar con un modelo de IA para tamizar los estudios que requieren informe de manera automática reduciría significativamente la carga de trabajo de los profesionales, permitiéndoles dedicar ese tiempo a por ejemplo tratar la patología de los pacientes. Esto sólo es posible, sin embargo, si el algoritmo asegura resultados precisos, tanto en términos de su seguridad (es decir, que no ignore casos de riesgo tratándolos como casos negativos) como de su eficiencia (que no tenga un gran número de falsos positivos). De la evaluación objeto de este trabajo, realizada sobre más de 61.000 estudios, se desprende que el modelo de IA considerado es capaz de obtener valores de AUC superiores a 0.95 para el reconocimiento de RD referible. Al individualizar el análisis para cada subconjunto de estudio (Tabla 1), se observa que estos valores son mayormente estables, llegando en algunos casos a un AUC = 1. Esto puede entenderse como un signo de su robustez a cambios en el dispositivo de captura y a variaciones en la población de personas estudiadas.

En lo que respecta a la SE y ESP del modelo, se observa que para los puntos operativos PO1 y PO2 es posible obtener resultados que o bien maximizan la capacidad de tamizaje de casos de riesgo (PO1) o bien minimizan la carga de trabajo excedente producto de la presencia de demasiados falsos positivos (PO2). El PO50% reporta resultados muy similares a los del PO2, ya que el PO2 trabaja con una probabilidad umbral de 51.23%, muy cercana al 50% empleado por el PO50%. Esto sugiere que las probabilidades de salida del modelo están calibradas hacia mayores valores de ESP, consecuencia de que fue entrenado con una proporción mayor de casos no referibles que de referibles. Esto puede resolverse reentrenando al algoritmo para que penalice más los errores cometidos para la clase referible y recalibre las probabilidades de salida. Se destaca, sin embargo, la efectividad del modelo para el PO1, que permite identificar a la gran mayoría de los casos de riesgo (95.5%) para un 30% de falsos positivos. Esto implica contar con un sistema capaz de reconocer a toda persona con RD referible y que alivia en un 70% la carga de trabajo del nodo de informes para falsos positivos. Esto permite hacer un mejor uso del recurso humano disponible.

Respecto al análisis cualitativo de las respuestas del algoritmo, se observó 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 imágenes en las que no es posible observar la mácula (por ejemplo, las centradas en la papila o en regiones periféricas), se observó que el algoritmo se enfoca en el estado del disco óptico y las arcadas. Los mapas también sugieren que las predicciones de RD referible se sustentan en detectar lesiones hemorrágicas y exudados blandos algodonosos y/o duros en la cercanía del área macular. Esto es compatible con algunos de los criterios aplicados en la práctica médica, donde no solo son tenidos en cuenta como marcadores diagnósticos sino también, por su localización, como predictores de amenaza de la función visual. Estos criterios no fueron modelados explícitamente al diseñar el algoritmo, sino que fueron aprendidos por éste a partir de los datos de entrenamiento, por lo que es de esperar que incorporarlos explícitamente pueda fortalecer aún más los resultados. En escenarios en los que los pacientes presentan condiciones diferentes a la RD, el algoritmo puede predecir RD referible si observa estructuras o lesiones 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 como falsos positivos a aquellas imágenes con otras condiciones relevantes, que aunque no son RD deben ser analizadas por un profesional. No obstante, queda pendiente realizar una evaluación más exhaustiva, por ejemplo cuantificando su habilidad diagnóstica en casos con otros enfermedades y etiquetados específicamente.

Por otro lado, se observó que la retroalimentación que el modelo puede ofrecer al profesional mediante los mapas de activación de clase no es tan precisa como la que pueden otros modelos orientados explícitamente en segmentar lesiones [9], que las delinean con precisión y permiten utilizar sus máscaras binarias para obtener estadísticas sobre su ubicación respecto al disco óptico o la mácula, o el porcentaje de la imagen ocupado por ellas. Dichos enfoques, sin embargo, requieren entrenarse con imágenes marcadas píxel a píxel por un profesional, que son difíciles de obtener. Además, suelen no ser robustos para el tamizado de casos de riesgo de RD. En un futuro el modelo evaluado podría mejorarse incorporando soporte para que prediga también la presencia/ausencia de neovascularizaciones, hemorragias, exudados, etc, y la generación de mapas de activación explícitos para cada una de esas predicciones. De esta forma, los mismos podrían ser más localizados y retroalimentar de mejor manera al operador de la herramienta.

Finalmente, un último punto a señalar tiene que ver con los resultados obtenidos sobre imágenes nacionales. Aunque el número de estudios utilizados es inferior al de los demás conjuntos, los valores de AUC, SE y ESP reportados en ellas son compatibles con los demás, 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-midriática no utilizado para el entrenamiento del algoritmo.

4. Conclusiones

En este trabajo realizamos una evaluación retrospectiva de un algoritmo de IA nacional para el tamizaje automático de casos de RD referible basado en redes neuronales convolucionales. 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. Se observó que el modelo es capaz de lograr una tasa de detección de casos positivos superior al 95% para un 30% de falsos positivos. Esto sugiere un punto de partida aceptable para el desarrollo de un modelo predictivo nacional para satisfacer la demanda de tamizaje de Argentina. Para confirmar su eficacia, queda pendiente realizar un estudio retrospectivo a gran escala con un número mayor de imágenes adquiridas en nuestro país, y su aplicación prospectiva en un contexto controlado para estudiar su capacidad discriminativa en el marco de una campaña de tamizaje.

Referencias

[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] 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.

[4] 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.

[5] Pedregosa, F. Scikit-learn: Machine learning in Python. J Mach Learn Res, 2011; 12: 2825-2830.

[6] Wilson, EB. Probable inference, the law of succession, and statistical inference. J Am Stat Assoc, 1927; 22: 209-12

[7] 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.

[8] Fernandez, FG. TorchCAM: class activation explorer. Online: https://github. com/frgfm/torch-cam. Accedido por última vez el 27/3/2022.

[9] 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.

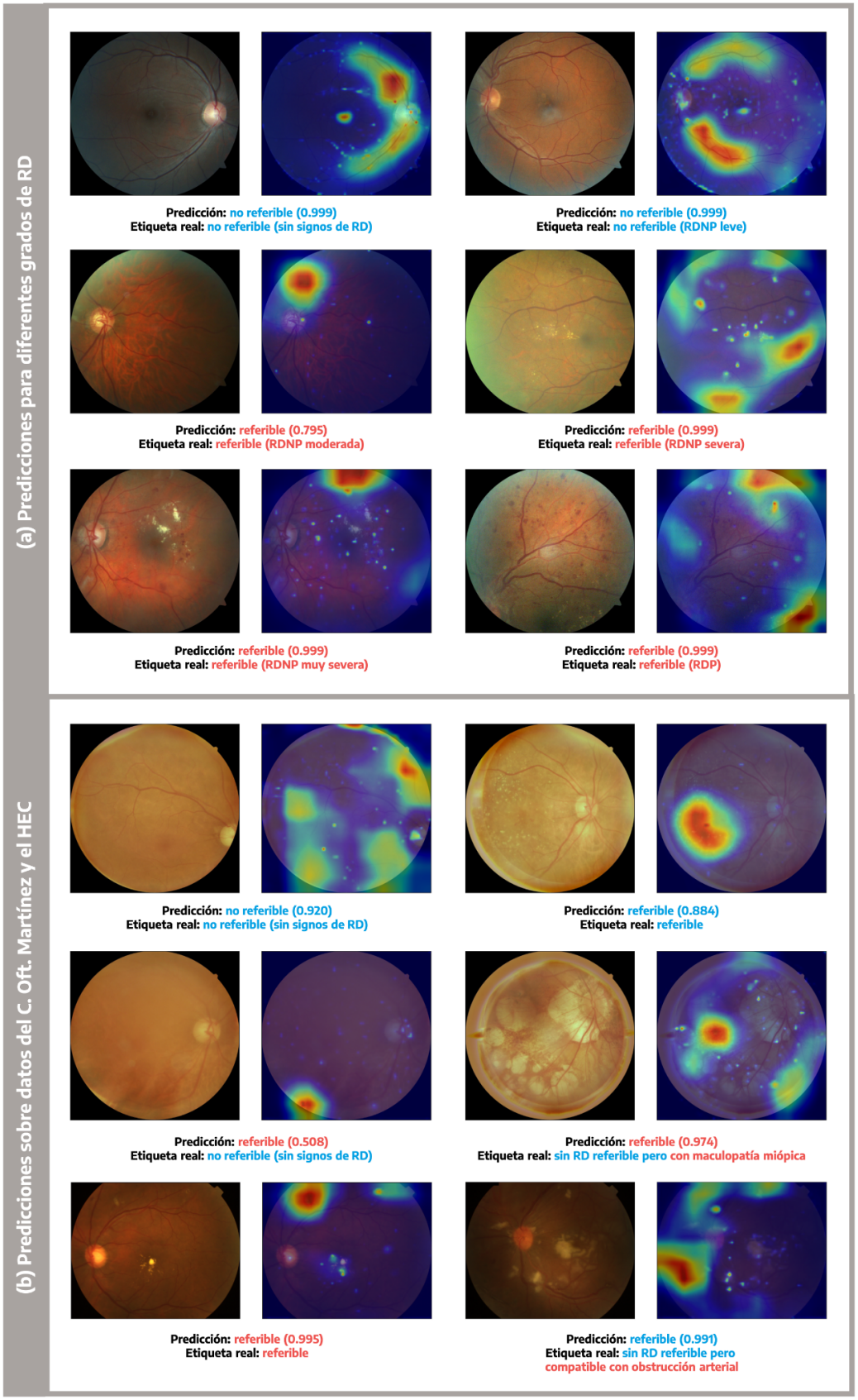
[10] Koch, et al. Estrategias innovadoras para mejorar los cuidados a personas con Enfermedades Crónicas. Reporte técnico, 2019. Ministerio de Salud de la Provincia de Buenos Aires.

[11] 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.

Tabla 1. Resultados obtenidos por el modelo propuesto sobre cada subconjunto de datos utilizado para evaluación, en términos de área bajo la curva ROC (AUC) y de sensibilidad (SE) y especificidad (ESP) para el punto operativo ubicado al considerar un umbral del 50% para predecir los casos como RD referible (PO50%). Se incluyen para cada métrica su correspondiente intervalo de confianza al 95% (IC 95%).

Conjunto	**AUC (IC 95%)**	**Para el PO50%**	
		SE (IC 95%)	**ESP (IC 95%)**
EyePACS (N = 53.576)	0.951 (0.949 - 0.954)	73.2% (72.3% - 74.0%)	97.9% (97.8% - 98.0%)
DDR (N = 3.759)	0.965 (0.960 - 0.970)	74.9% (72.8% - 76.9%)	97.8% (97.1% - 98.4%)
MESSIDOR 2 (N = 1.744)	0.973 (0.967 - 0.979)	89.5% (86.3% - 92.0%)	94.1% (92.7% - 95.3%)
FCM-UNA (N = 757)	0.986 (0.980 - 0.992)	88.2% (85.2% - 90.6%)	99.0% (96.3% - 99.7%)
DR2 (N = 435)	0.974 (0.962 - 0.985)	84.7% (76.3% - 90.5%)	96.1% (93.5% - 97.7%)
DeepDRID (N = 400)	0.959 (0.944 - 0.972)	88.3% (82.8% - 92.2%)	86.8% (81.7% - 90.7%)
1000Fundus (N = 144)	1.000 (1.000 - 1.000)	100% (95.8% - 100%)	96.4% (87.9% - 99.0%)
IDRiD (N = 103)	0.949 (0.914 - 0.980)	82.8% (71.8% - 90.1%)	89.7% (76.4% - 95.9%)
DIARETDB1 (N = 89)	0.981 (0.956 - 0.999)	95.7% (85.5% - 98.8%)	93.0% (81.4% - 97.6%)
Martínez (N = 484)	0.955 (0.927 - 0.980)	80.0% (62.7% - 90.5%)	93.4% (90.7% - 95.3%)
HEC (N = 35)	0.961 (0.900 - 1.000)	100% (61.0% - 100%)	86.2% (69.4% - 94.5%)

Figura 2. Algunos ejemplos de resultados cualitativos generados por el modelo de inteligencia artificial. (a) Predicciones correctas para diferentes grados de RD observada del conjunto FCM-UNA. (b) Predicciones sobre imágenes de los bancos de datos Martínez y HEC.


(a) Predicciones para diferentes grados de RD
Predicción: no referible (0.999)
Etiqueta real: no referible (sin signos de RD)
Predicción: no referible (0.999)
Etiqueta real: no referible (RDNP leve)
Predicción: referible (0.795)
Etiqueta real: referible (RDNP moderada)
Predicción: referible (0.999)
Etiqueta real: referible (RDNP severa)
Predicción: referible (0.999)
Etiqueta real: referible (RDNP muy severa)
Predicción: referible (0.999)
Etiqueta real: referible (RDP)
(b) Predicciones sobre datos del C. Oft. Martínez y el HEC
Predicción: no referible (0.920)
Etiqueta real: no referible (sin signos de RD)
Predicción: referible (0.884)
Etiqueta real: referible
Predicción: referible (0.508)
Etiqueta real: no referible (sin signos de RD)
Predicción: referible (0.974)
Etiqueta real: sin RD referible pero con maculopatía miópica
Predicción: referible (0.995)
Etiqueta real: referible
Predicción: referible (0.991)
Etiqueta real: sin RD referible pero
compatible con obstrucción arterial