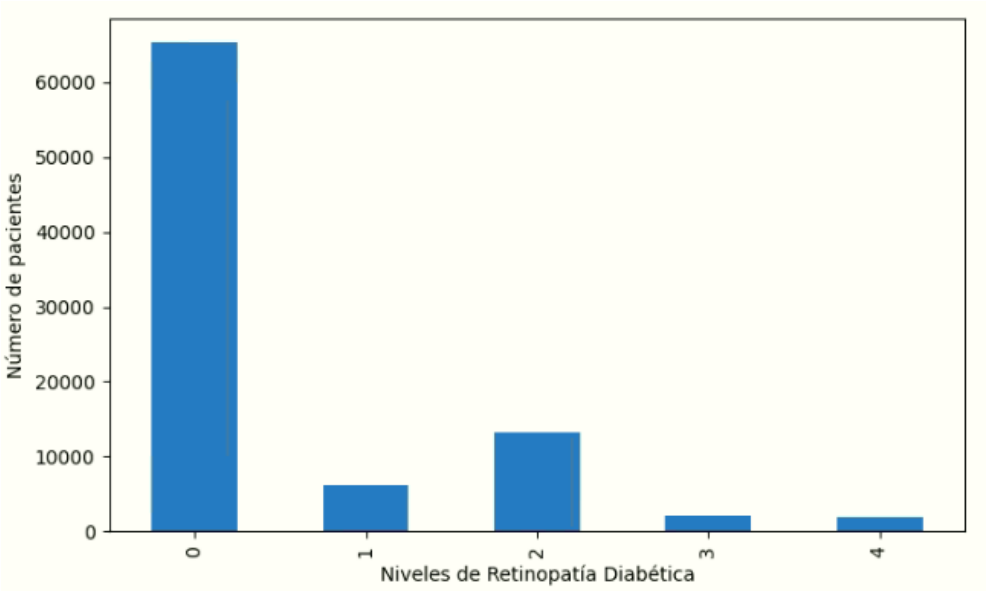
Detección de Retinopatía Diabética en imágenes de fondo de ojos

Con el objetivo de crear una herramienta capaz de clasificar de forma autónoma las retinografías en función de la presencia o no de Retinopatía Diabética (RD), se ha entrenado una red neuronal convolucional con miles de imágenes de este tipo.

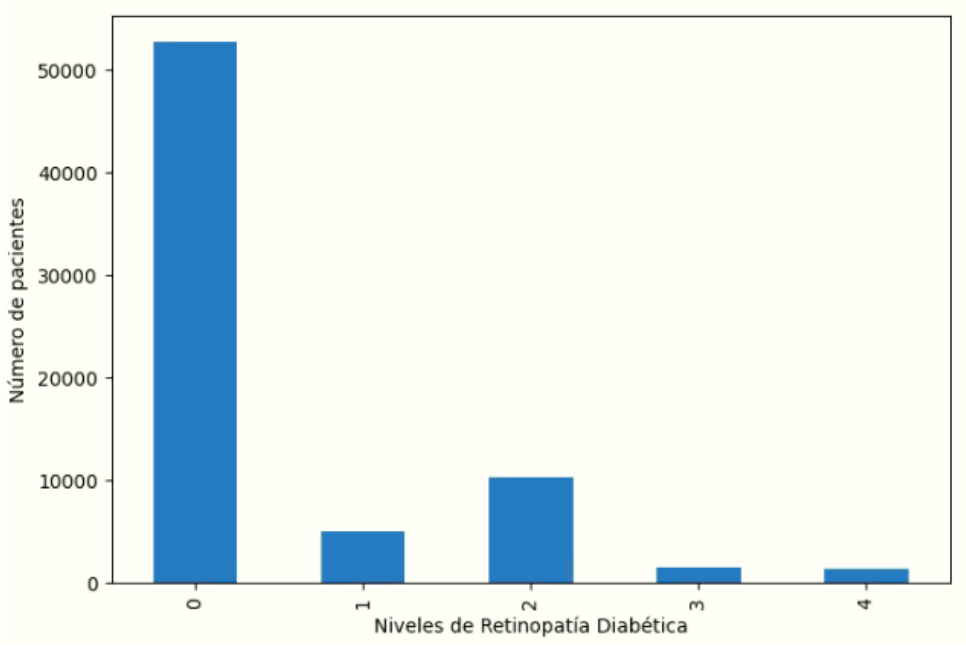
Datos empleados

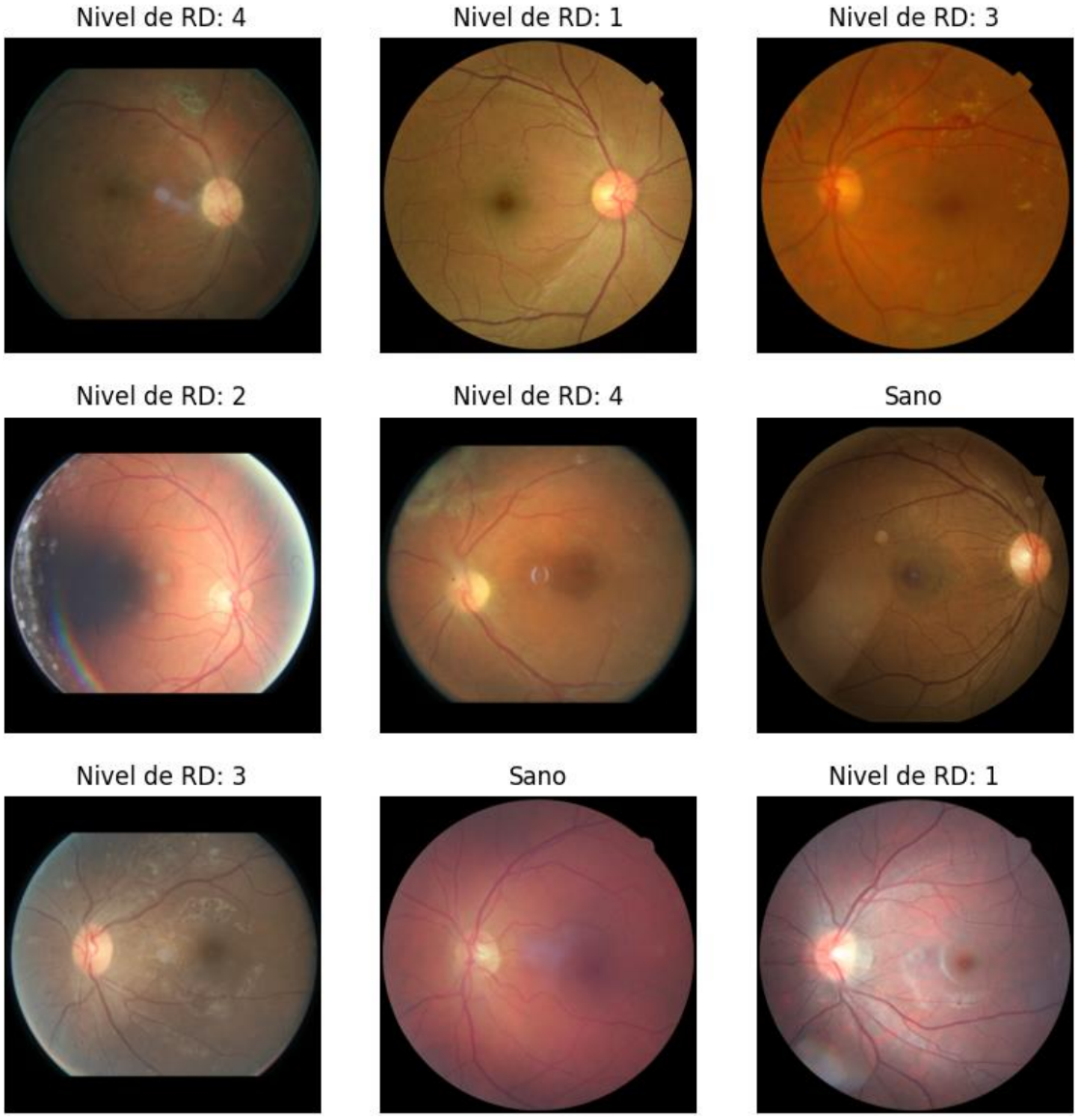
El conjunto de imágenes que se ha empleado para el entrenamiento de esta red se encuentra públicamente disponible en la web de Kaggle disponible en [1]. En total, se compone de 88.602 imágenes provenientes de EYEPACS, clasificadas por un especialista (oftalmólogo) para los 5 niveles de RD, siendo esta su distribución:



Es claramente apreciable un desequilibrio entre el número de sanos y el número de pacientes con la enfermedad. Esto será convenientemente tratado más adelante.

Sin embargo, tal y como se describe en la página, algunas de las imágenes pueden contener ruidos, artefactos, distorsiones… elementos visuales que podrían dificultar o imposibilitar un correcto diagnóstico. Por tanto, y a partir del etiquetado realizado por uno de los autores de este trabajo [2], se han descartado un total de 17.641 imágenes cuya calidad era insuficiente para diagnosticar, quedando un total de 71.051. La nueva distribución del conjunto de datos, una vez apartadas dichas imágenes, queda muy similar a la anterior:





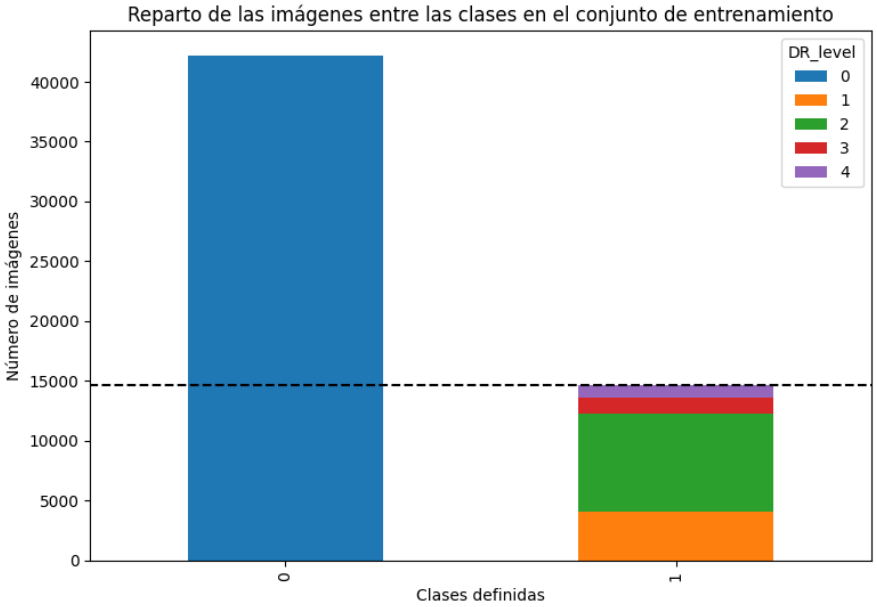
Muestra de imágenes pertenecientes a los 5 niveles de RD

Posteriormente, estas imágenes han sido repartidas aleatoriamente en 3 subconjuntos de distinto tamaño:

- Un 80% del total (56.822) constituirá el conjunto de entrenamiento. Ninguna imagen ajena a este conjunto será utilizada para ajustar la red.
- Otro 10% (7.103) formará el conjunto de validación.
- Y el 10% restante (7.102) dará forma al conjunto de test/prueba.

La clasificación final deseada será diferenciar todos aquellos pacientes que padezcan de RD en cualquiera de sus niveles (mayores que 0). En consecuencia, todas las imágenes correspondientes a los grados 1 o mayores, serán agrupados como una única clase.

No obstante, al haber realizado la división de forma aleatoria, los tres nuevos conjuntos aún poseen una distribución desigual en la que predominan los sanos frente a los demás. Este hecho podría condicionar fuertemente el entrenamiento de la red y, por ende, los resultados que pudiera obtener.



La imagen superior muestra la desbalanceada proporción entre sanos y demás niveles (casi 3:1) en el conjunto destinado al entrenamiento de la red convolucional. Concretamente, el número de imágenes pertenecientes a cada nivel de RD es:

- Con nivel 0 de RD (sanos): 42.149 imágenes
- Con nivel 1: 4.049 imágenes
- Con nivel 2: 8.241 imágenes
- Con nivel 3: 1.283 imágenes
- Con nivel 4: 1.100 imágenes

Para resolver este desequilibrio en la etapa de aprendizaje, sólo 14.673 imágenes de la clase de sanos serán analizadas, haciendo un total de 29.346 imágenes contando los pacientes con la enfermedad. De esta forma, para cada clase definida (sanos y con enfermedad, sin distinciones directas entre los niveles de RD), siempre habrá el mismo número de muestras, solucionando el posible sesgo.

Durante el entrenamiento, se alimenta a la red con todas las imágenes de dicho conjunto. Una vez son analizadas todas las imágenes de entrenamiento, se dice que ha completado una época.

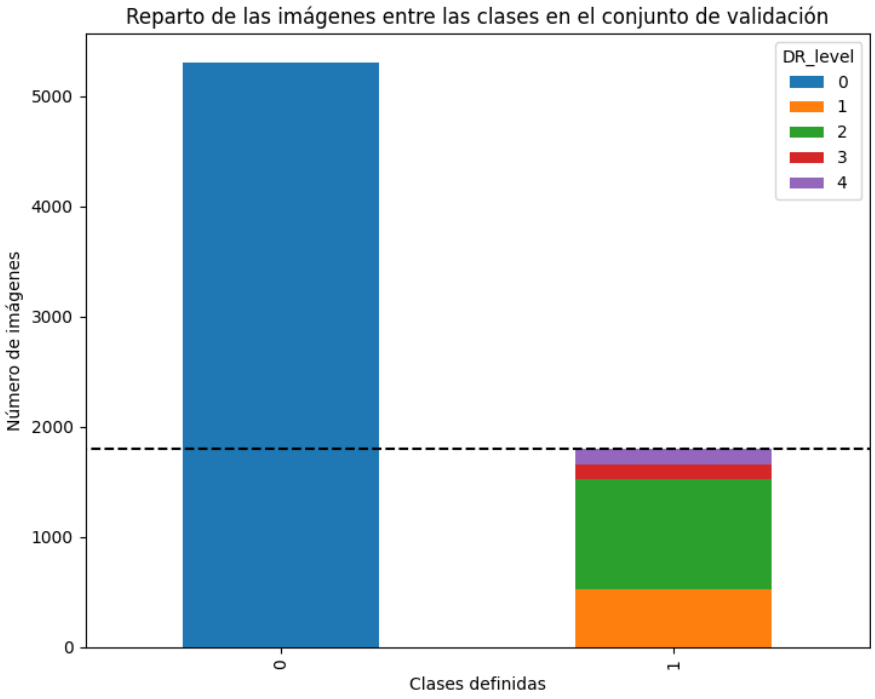
Época a época la red irá mejorándose de forma que sus predicciones sean cada vez más acertadas.

Con idea de no desaprovechar las casi 28.000 imágenes descartadas, al inicio de cada época se tomarán 14.673 imágenes aleatorias de todas las pertenecientes a la clase de sanos. De esta forma, se evita tomar las mismas imágenes de pacientes sanos continuamente.

Un proceso similar será aplicado al conjunto de validación. Este conjunto posee una distribución casi idéntica donde, aproximadamente, por cada imagen perteneciente a la clase de pacientes con la enfermedad hay 3 imágenes de sanos. Específicamente, la cantidad de imágenes pertenecientes a cada nivel de RD es la siguiente:

- Con nivel 0: 5.304 imágenes
- Con nivel 1: 522 imágenes
- Con nivel 2: 997 imágenes
- Con nivel 3: 138 imágenes
- Con nivel 4: 142 imágenes

Siendo su distribución la que se muestra a continuación:



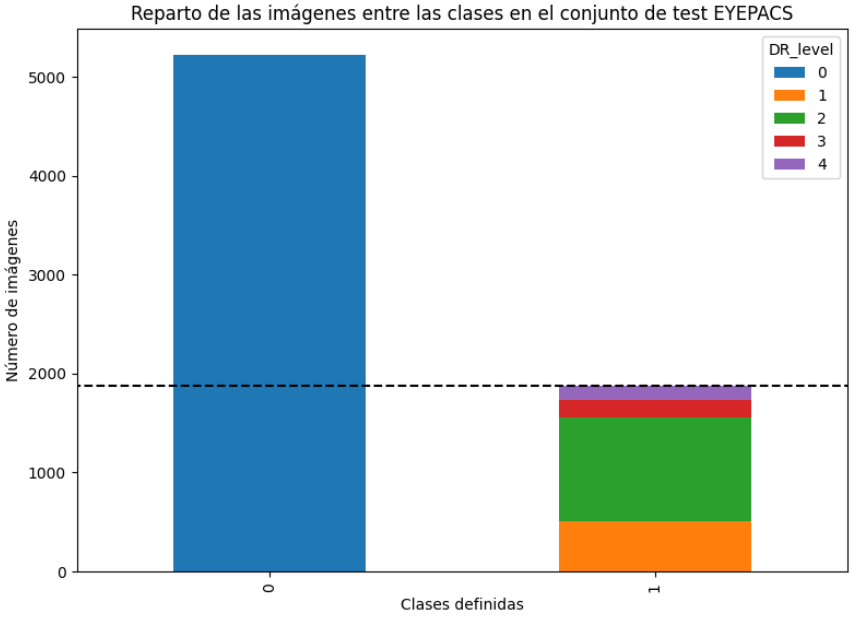
Al igual que en el conjunto de entrenamiento, se tomarán exactamente el mismo número de imágenes de cada clase: 1.799, con la excepción de que, en este conjunto, las imágenes de la clase 0 no variarán durante el entrenamiento y se tomarán siempre las mismas. Esto se debe a que, cada vez que finaliza una época de entrenamiento, la red es sometida a una evaluación sobre un conjunto aparte de imágenes (el conjunto de validación). Mediante esta acción, se podrá llevar a cabo un control más preciso del rendimiento real de la red. Es por ello por lo que no es deseable que este conjunto pueda variar sus elementos conforme el entrenamiento avanza. El número total de imágenes de validación será, por tanto, de 3.598, que no se verán modificadas en ningún momento del entrenamiento.

Por el otro lado, el conjunto de test/prueba permanecerá estático todo el tiempo. La proporción entre sanos y pacientes enfermos se mantendrá tal cual existe en el conjunto de imágenes

original. Esto se debe a que la proporción presente en los datos es prácticamente igual a la proporción existente en la vida real. En vista de ello, evaluar el modelo (tras finalizar su entrenamiento) sobre este conjunto de imágenes proporcionará datos más fiables acerca de cómo funcionaría esta red en caso de comenzar a utilizarse como herramienta de cribado.

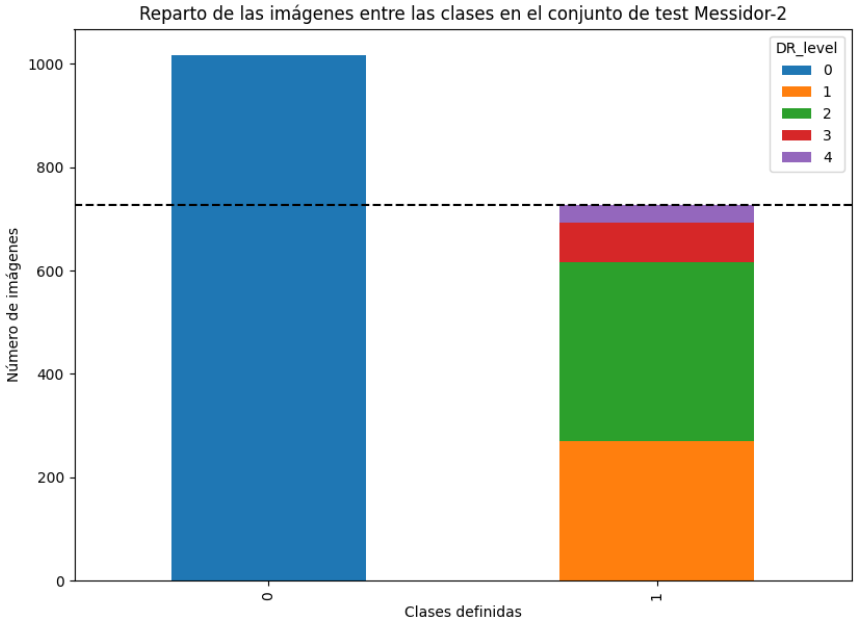
Esto hace que la distribución de las imágenes en el conjunto de test sea la siguiente:

- Con nivel 0: 5.223 imágenes
- Con nivel 1: 505 imágenes
- Con nivel 2: 1.055 imágenes
- Con nivel 3: 176 imágenes
- Con nivel 4: 143 imágenes



Además, haciendo énfasis en la medición del rendimiento de la red sobre nuevas imágenes, se hará uso también del conjunto de imágenes Messidor-2, disponible en [3]. Este conjunto se compone de 1.748 imágenes no etiquetadas, sin embargo, se hará uso del etiquetado disponible en Kaggle [4] proporcionado por los autores de [5]. Este etiquetado, además, fue asignado mediante el consenso entre tres oftalmólogos. De todas las imágenes, 4 serán descartadas por no poseer ningún etiquetado debido a su carencia de calidad suficiente para un diagnostico correcto. La distribución queda de la siguiente manera:

- Con nivel 0: 1.017 imágenes
- Con nivel 1: 270 imágenes
- Con nivel 2: 347 imágenes
- Con nivel 3: 75 imágenes
- Con nivel 4: 35 imágenes

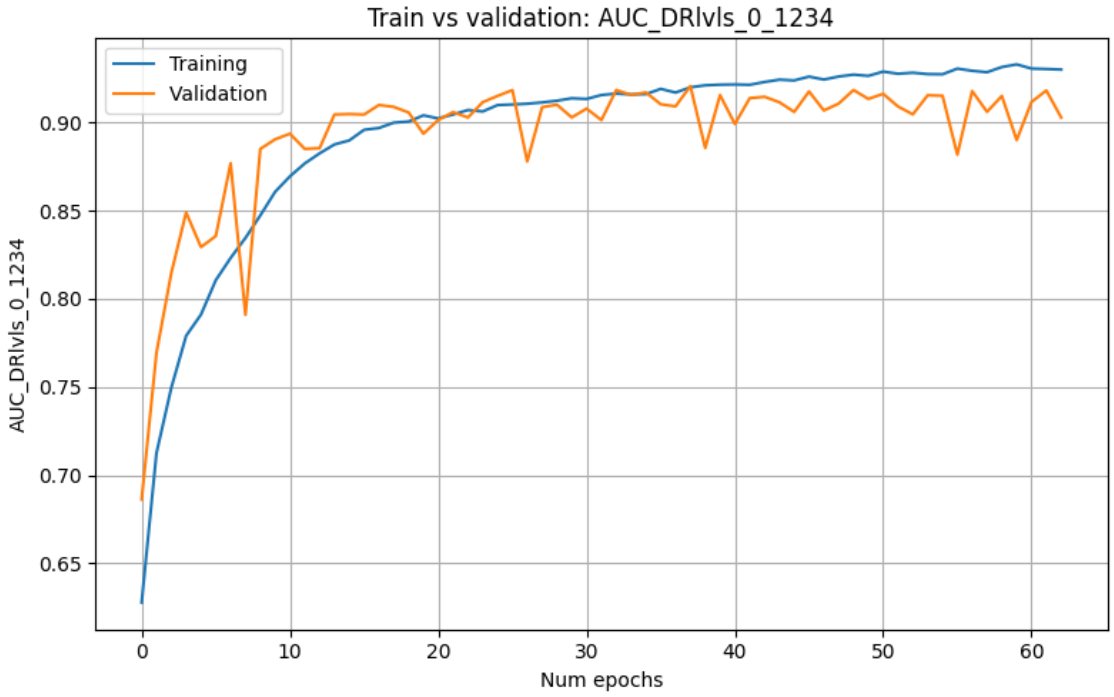


Nuevamente, la presencia de imágenes correspondientes a pacientes sanos es mayor que las demás clases, aunque en este conjunto, la diferencia no es tan marcada.

Resultados obtenidos

Durante el entrenamiento, tras cada época, la red fue evaluada sobre el conjunto de validación, recogiendo sus resultados. De las métricas empleadas, se hará un mayor hincapié en el AUC (Area Under the ROC) obtenido en validación para la clasificación entre las clases 0 y 1, es decir, sanos y pacientes con la enfermedad.

Una vez se da por finalizado el entrenamiento, se procede a evaluar la red sobre los conjuntos de testeo descritos previamente. Concretamente, la red a examinar será la que obtuvo un mejor valor de AUC en validación durante todo el entrenamiento. Esto sucedió en la época 37 del entrenamiento.



Primero, se obtienen los resultados que logra la red seleccionada sobre el conjunto de validación. Esto permitirá seleccionar diferentes umbrales que garanticen un cierto nivel de sensibilidad en la predicción de casos de enfermedad.

Los umbrales seleccionados serán aquellos que garanticen al menos un 98, 95, 93, 90 y 85% de sensibilidad en la predicción sobre el conjunto de validación. Además, se agregará un sexto umbral que se corresponda al punto más cerca de la curva ROC al punto (0,1), es decir, máxima sensibilidad y especificidad.

Resultados en el conjunto de validación

El valor de AUC obtenido ha sido de: 0.924

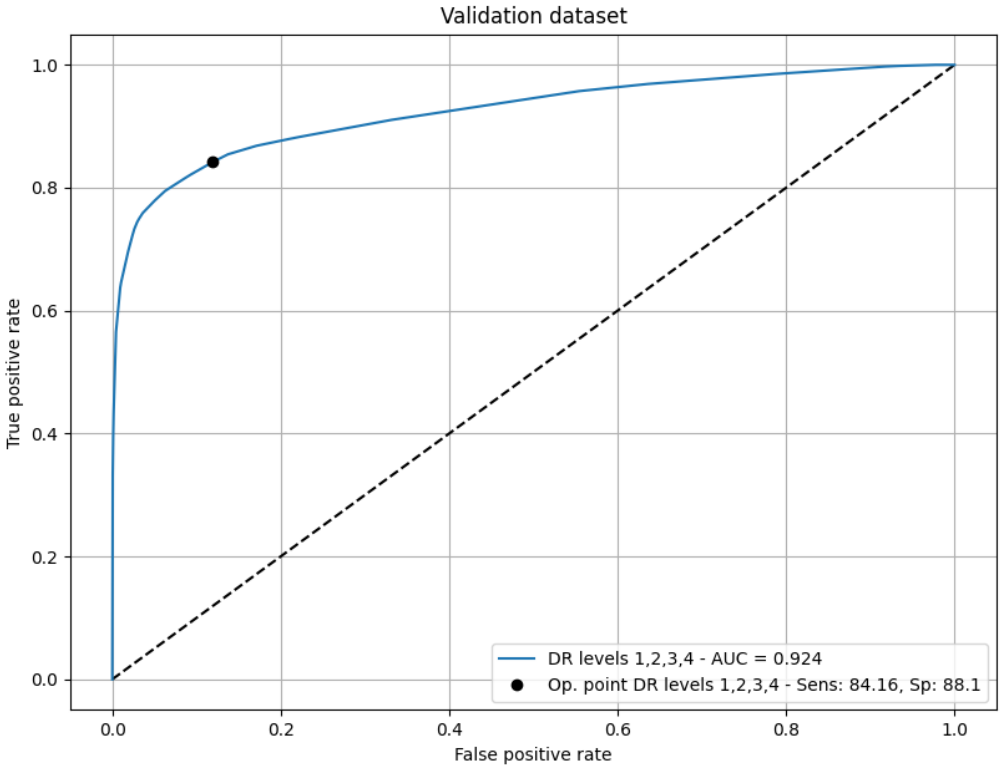
Punto de operación	Sensibilidad	Especificidad	Tasa de Falsos Negativos (TFN)	TFN Nivel 1	TFN Nivel 2	TFN Nivel 3	TFN Nivel 4	Umbrales
98	98.50	21.40	1.50	3.07	1.10	0.00	0.00	0.060120
95	95.72	44.58	4.28	9.00	2.91	0.00	0.70	0.108216
93	95.72	44.58	4.28	9.00	2.91	0.00	0.70	0.108216
90	91.05	66.81	8.95	18.97	5.82	0.72	2.11	0.190381
85	85.44	86.27	14.56	32.38	8.43	1.45	4.93	0.368737
Punto más cercano	84.16	88.10	15.84	35.06	9.33	1.45	4.93	0.410822

Se ha marcado aquella fila cuyo umbral garantiza al menos un 90% de sensibilidad y al menos un 60% de especificidad.

Con este umbral, el número exacto de falsos negativos por cada nivel de RD ha sido:

- Para el nivel 1: 99 de los 522 existentes han sido marcados como negativos
- Para el nivel 2: 58 de los 997 casos.
- Para el nivel 3: 1 de los 138 casos.
- Para el nivel 4: 3 de los 142 casos.

Empleando el umbral señalado, un 78.93% de las imágenes fueron correctamente clasificadas. Por otra parte, si se emplea el umbral que más cerca queda del punto (0,1), el número de imágenes correctamente clasificadas asciende al 86.13%. Esto se debe a la alta diferencia en especificidad entre ambos umbrales.



A continuación, se muestran una serie de imágenes de este conjunto que han sido catalogadas como sanas a pesar de no serlo:

Nivel real de RD: 4

Nivel real de RD: 4

Nivel real de RD: 4

Nivel real de RD: 3

Nivel real de RD: 2

Nivel real de RD: 2

Nivel real de RD: 1

Nivel real de RD: 1

Resultados en el conjunto de testeo de EYEPACS

El valor de AUC obtenido ha sido de: 0.9212

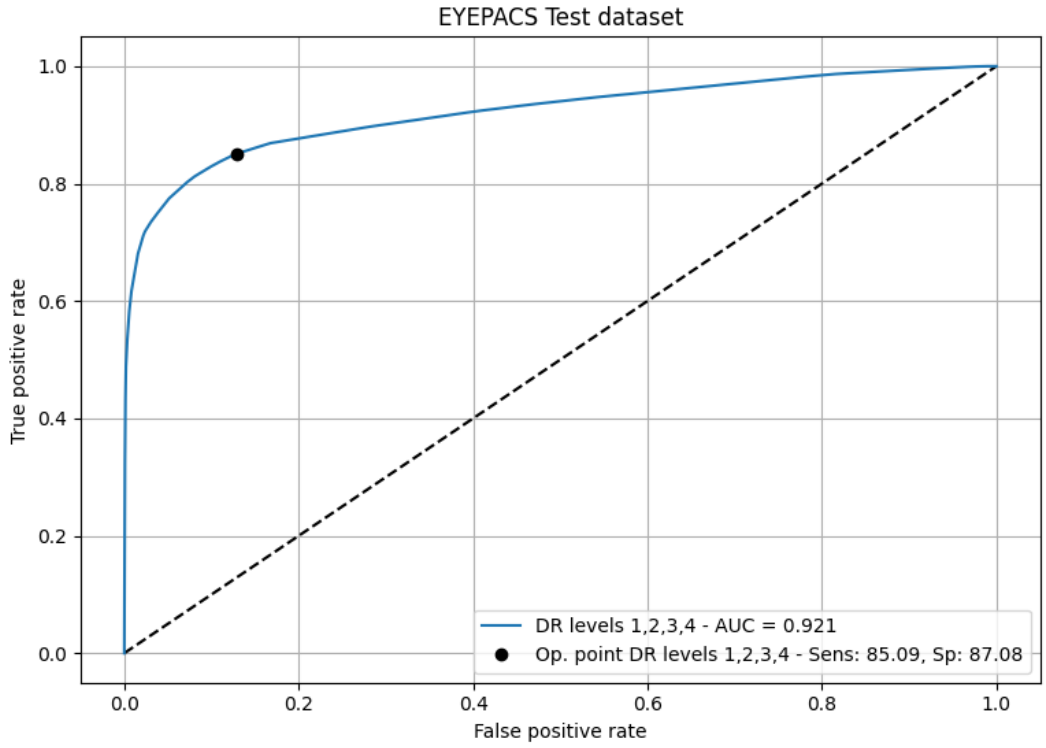
Punto de operación	Sensibilidad	Especificidad	Tasa de Falsos Negativos (TFN)	TFN Nivel 1	TFN Nivel 2	TFN Nivel 3	TFN Nivel 4	Umbrales
98	98.39	20.66	1.61	4.16	0.95	0.00	0.00	0.060120
95	95.36	41.48	4.64	11.29	3.41	0.57	0.00	0.108216
93	95.36	41.48	4.64	11.29	3.41	0.57	0.00	0.108216
90	91.25	64.66	8.75	20.20	6.54	0.57	0.70	0.190381
85	86.19	84.75	13.81	30.30	9.76	1.70	2.10	0.368737
Punto más cercano	85.09	87.08	14.91	32.28	10.71	1.70	2.80	0.410822

Los umbrales empleados han sido los que se obtuvieron al analizar los resultados en validación. Nuevamente, se ha marcado la fila cuyo umbral satisfizo la condición de un 90% de sensibilidad y un 60% de especificidad en el conjunto de validación. Como puede verse, esto se repite también en este conjunto.

Con este umbral, el número exacto de falsos negativos por cada nivel de RD ha sido:

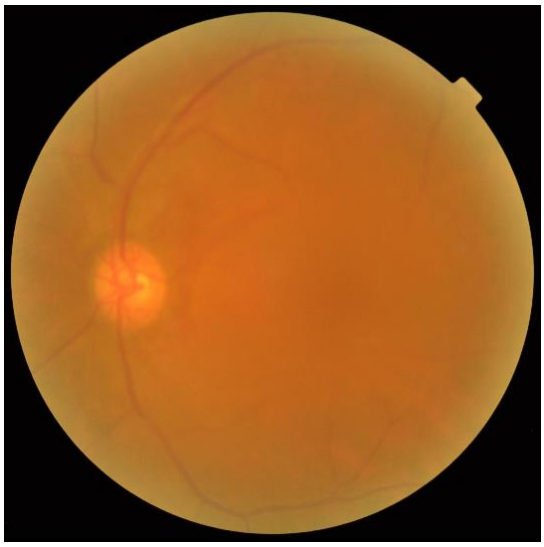
- Para el nivel 1: 102 de los 505 existentes han sido marcados como negativos
- Para el nivel 2: 69 de los 1.055 casos.
- Para el nivel 3: 1 de los 176 casos.
- Para el nivel 4: 1 de los 143 casos.

Haciendo uso del umbral señalado, la precisión de la predicción en este conjunto es de un 72.18%. En caso de aplicar el umbral más cercano a (0,1) en validación, ésta asciende hasta el 86.52%.

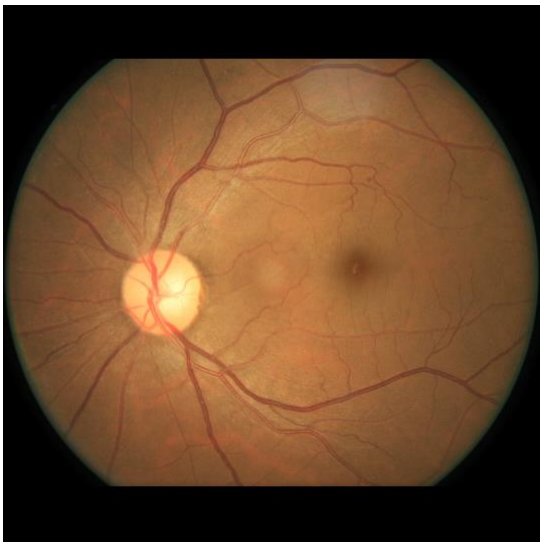


A continuación, se muestran una serie de imágenes que han sido catalogadas como sanas a pesar de no serlo:

Nivel real de RD: 4

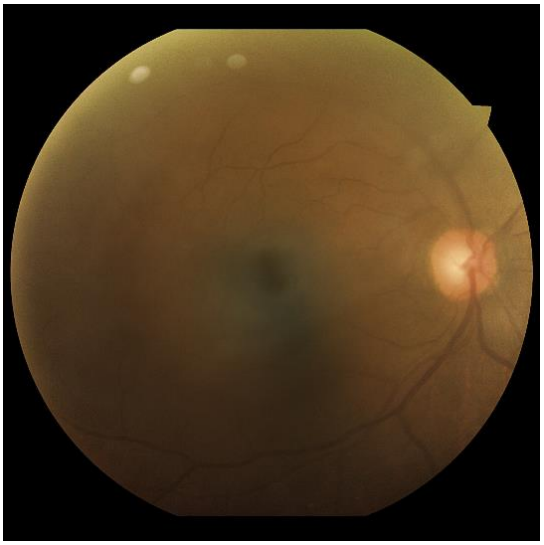


Nivel real de RD: 3



Nivel real de RD: 2

Nivel real de RD: 2



Nivel real de RD: 2

Nivel real de RD: 1

Nivel real de RD: 1

Nivel real de RD: 1

Resultados en el conjunto de testeo de Messidor-2

El valor de AUC obtenido ha sido de: 0.9303

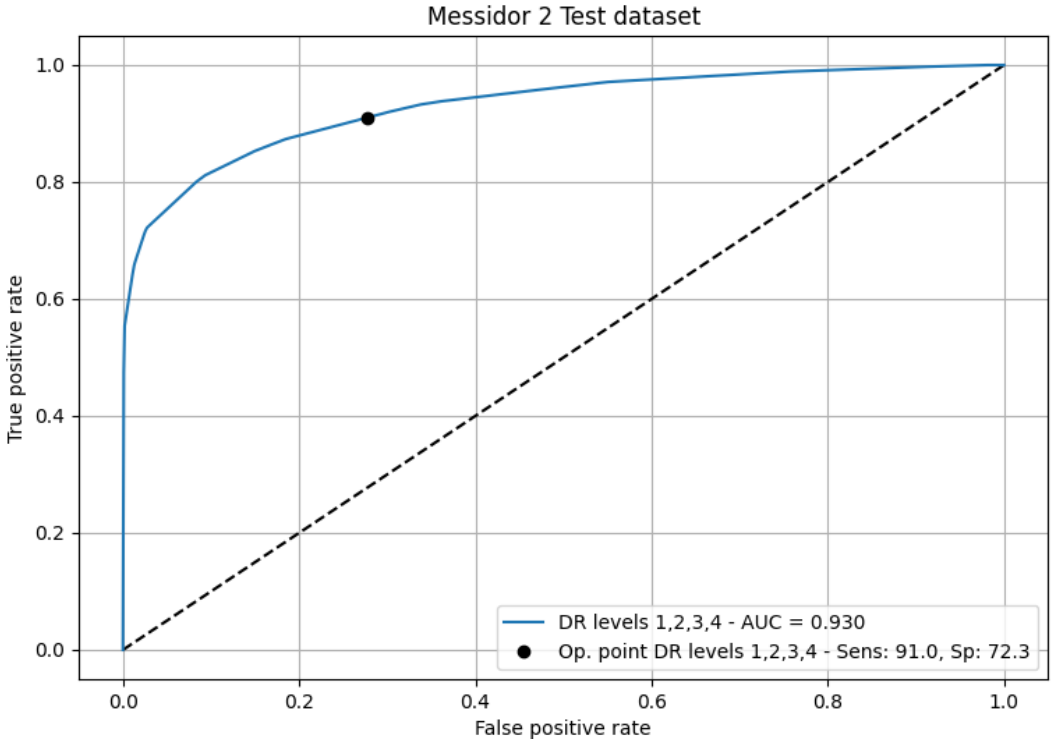
Punto de operación	Sensibilidad	Especificidad	Tasa de Falsos Negativos (TFN)	TFN Nivel 1	TFN Nivel 2	TFN Nivel 3	TFN Nivel 4	Umbrales
98	99.13	19.82	0.87	2.59	0.00	0.00	0.00	0.060120
95	97.84	36.51	2.16	6.67	0.00	0.00	0.00	0.108216
93	97.84	36.51	2.16	6.67	0.00	0.00	0.00	0.108216
90	95.45	54.58	4.55	12.59	0.00	0.00	0.00	0.190381
85	91.86	70.13	8.14	21.48	0.58	0.00	0.00	0.368737
Punto más cercano	91.00	72.30	9.00	24.44	1.15	0.00	0.00	0.410822

Los umbrales empleados han sido los que se obtuvieron al analizar los resultados en validación. Nuevamente, se ha marcado la fila cuyo umbral satisfizo la condición de un 90% de sensibilidad y un 60% de especificidad en el conjunto de validación. En este caso, la sensibilidad es ligeramente mayor en contra de la especificidad, que cae varios puntos.

Con este umbral, el número exacto de falsos negativos por cada nivel de RD ha sido:

- Para el nivel 1: 34 de los 270 existentes han sido marcados como negativos
- Para el nivel 2: ninguno de los 347 casos.
- Para el nivel 3: ninguno de los 75 casos.
- Para el nivel 4: ninguno de los 35 casos.

Haciendo uso del umbral señalado, la precisión de la predicción en este conjunto es de un 71.16%. En caso de aplicar el umbral más cercano a (0,1) en validación, ésta asciende hasta el 80.01%.



A continuación, se muestran una serie de imágenes que han sido catalogadas como sanas a pesar de no serlo:

Nivel real de RD: 1

Nivel real de RD: 1

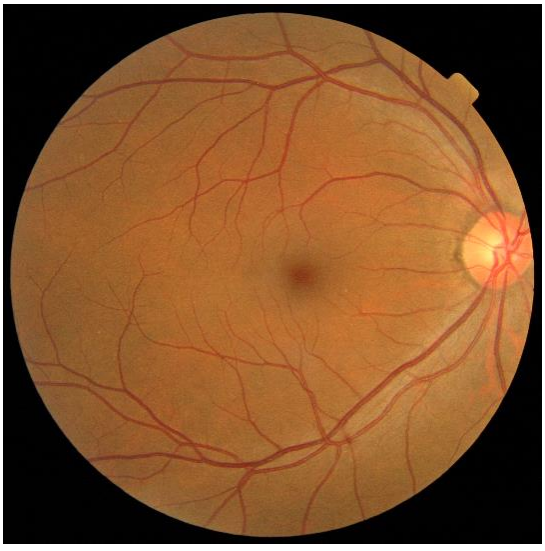
Nivel real de RD: 1

Nivel real de RD: 1

Nivel real de RD: 1

Nivel real de RD: 1

Nivel real de RD: 1



Nivel real de RD: 1

Referencias

[1] Kaggle, «Diabetic Retinopathy Detection (Data),» 2015. [En línea]. Available: https://www.kaggle.com/c/diabetic-retinopathy-detection/data.

[2] M. Voets, K. Møllersen y L. A. Bongo, «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,* vol. 14, nº 6, p. e0217541, 2019.

[3] «Messidor-2,» ADCIS, [En línea]. Available: https://www.adcis.net/en/third-party/messidor2/.

[4] «Messidor-2 DR Grades,» Kaggle, [En línea]. Available: https://www.kaggle.com/google-brain/messidor2-dr-grades.

[5] J. Krause, V. Gulshan, E. Rahimy, P. Karth, K. Widner, G. S. Corrado, ... y D. R. Webster, «Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy,» *Ophthalmology,* vol. 125, nº 8, pp. 1264-1272, 2018.