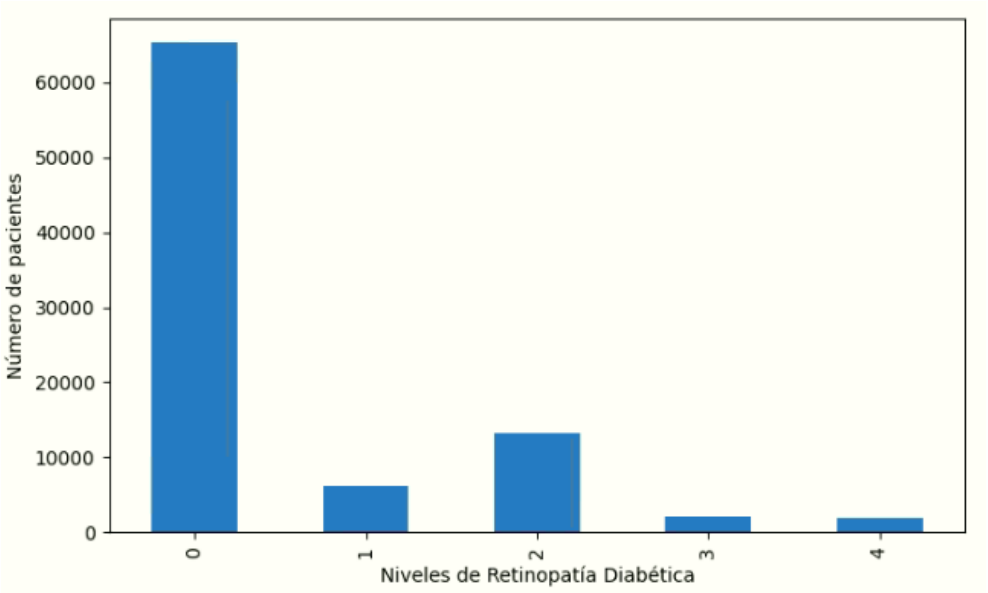
Detección de Retinopatía Diabética en diferentes niveles en imágenes de fondo de ojos

En este informe se presentarán los resultados obtenidos para la detección de Retinopatía Diabética en imágenes de fondo de ojo. Concretamente, se ha entrenado una red neuronal convolucional con miles de imágenes de este tipo con el objetivo de clasificarlas en tres clases: sanos, pacientes con los primeros indicios de la enfermedad (nivel 1 de ésta), y pacientes con signos evidentes de esta enfermedad (niveles 2, 3 y 4).

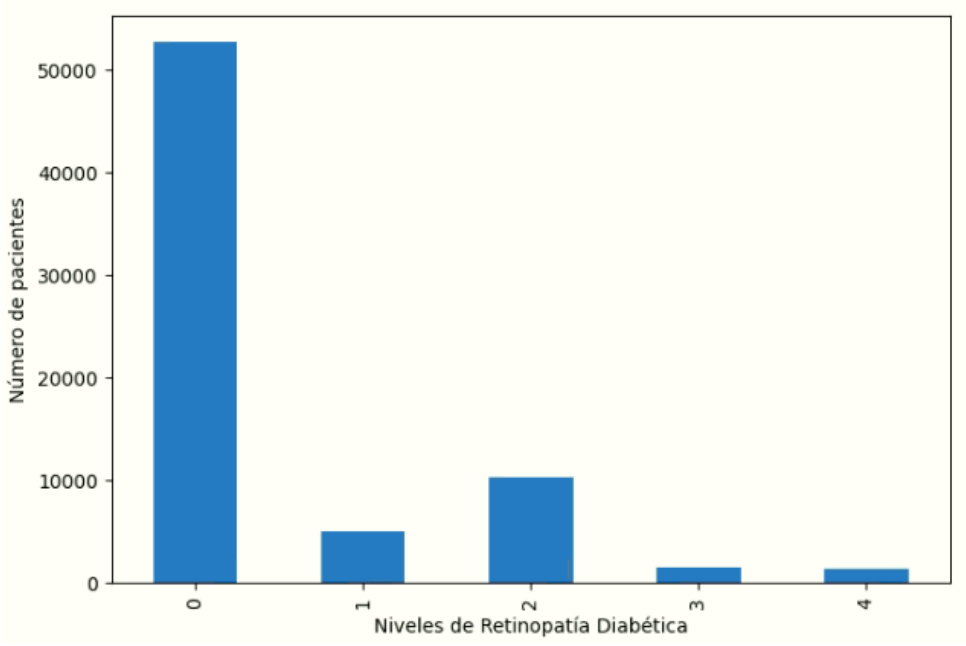
Datos empleados

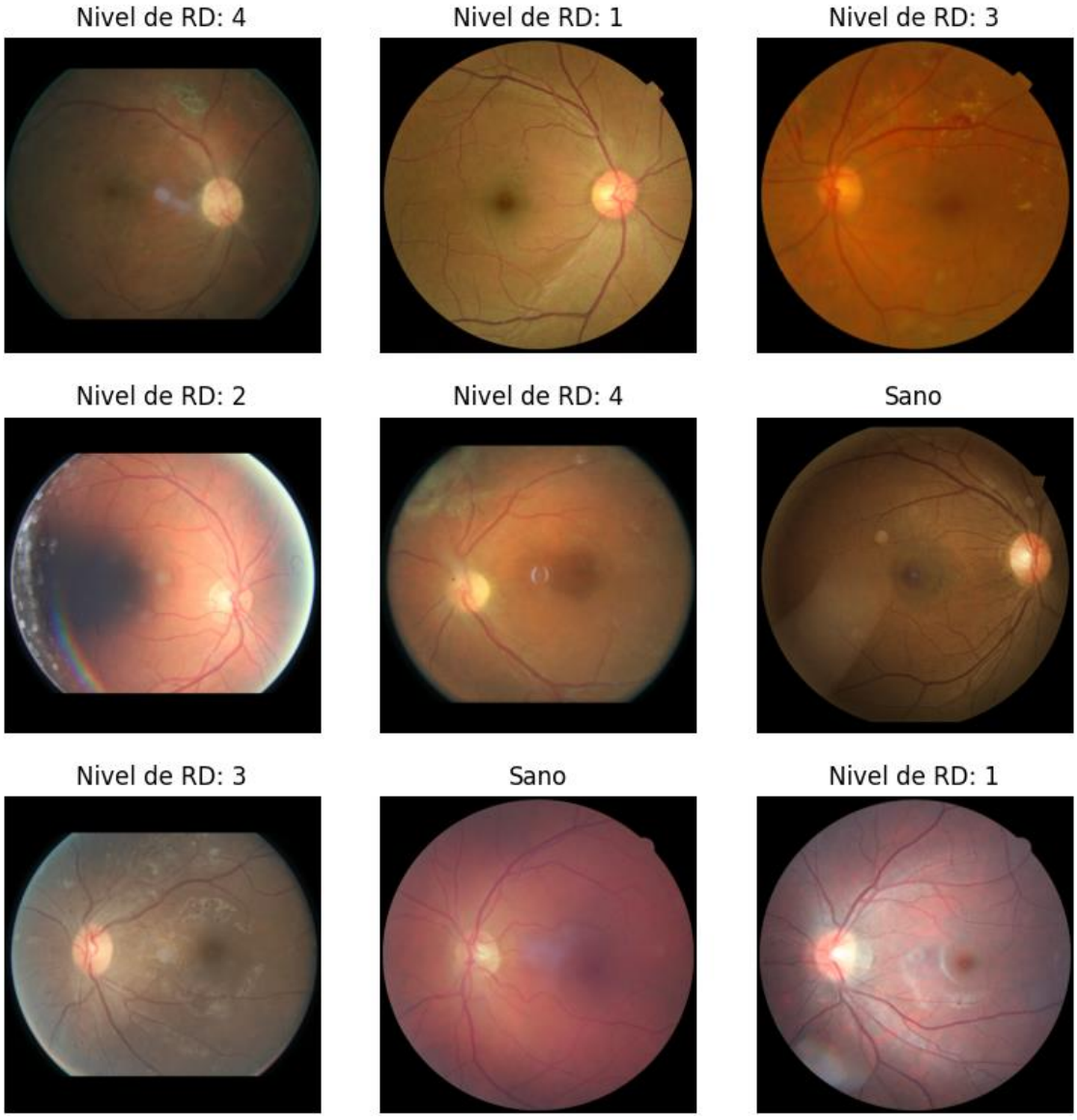
El conjunto de imágenes que se ha empleado para el entrenamiento de esta red se encuentra públicamente disponible en la web de Kaggle disponible en [1]. En total, se compone de 88.602 imágenes provenientes de EYEPACS, clasificadas por un especialista (oftalmólogo) para los 5 niveles de RD, siendo esta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





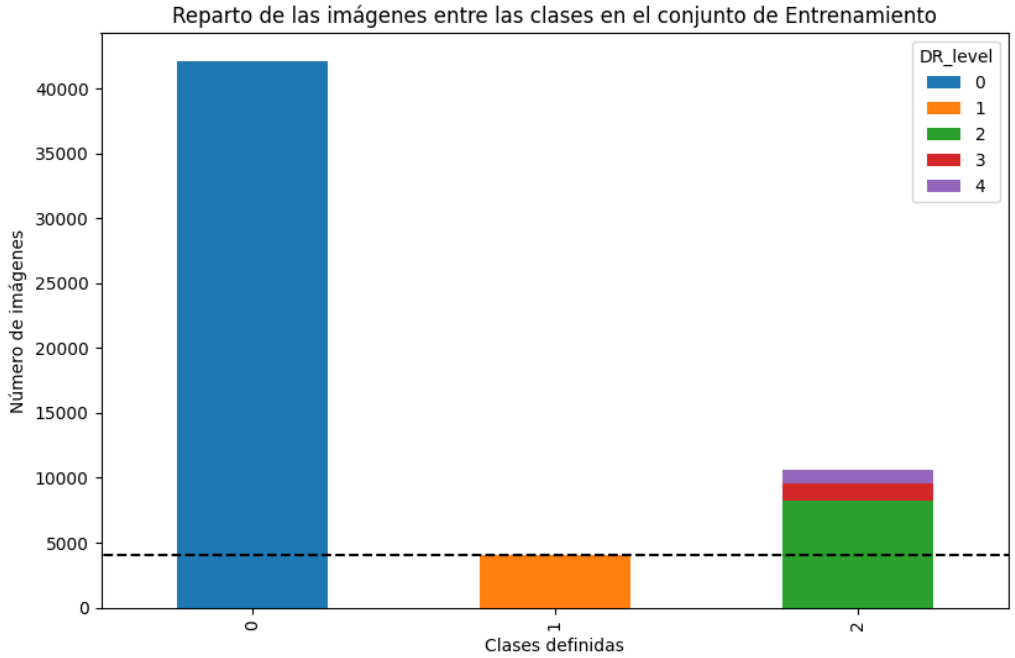
Muestra de imágenes pertenecientes a los 5 niveles de RD

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.

Estos conjuntos serán, a su vez, distribuidos en tres clases tal y como se describió al inicio: sanos, casos leves y casos moderados o peores, constituyendo las clases 0, 1 y 2.

No obstante, al haber realizado la división de forma aleatoria, los tres nuevos conjuntos aún poseen una distribución muy 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.

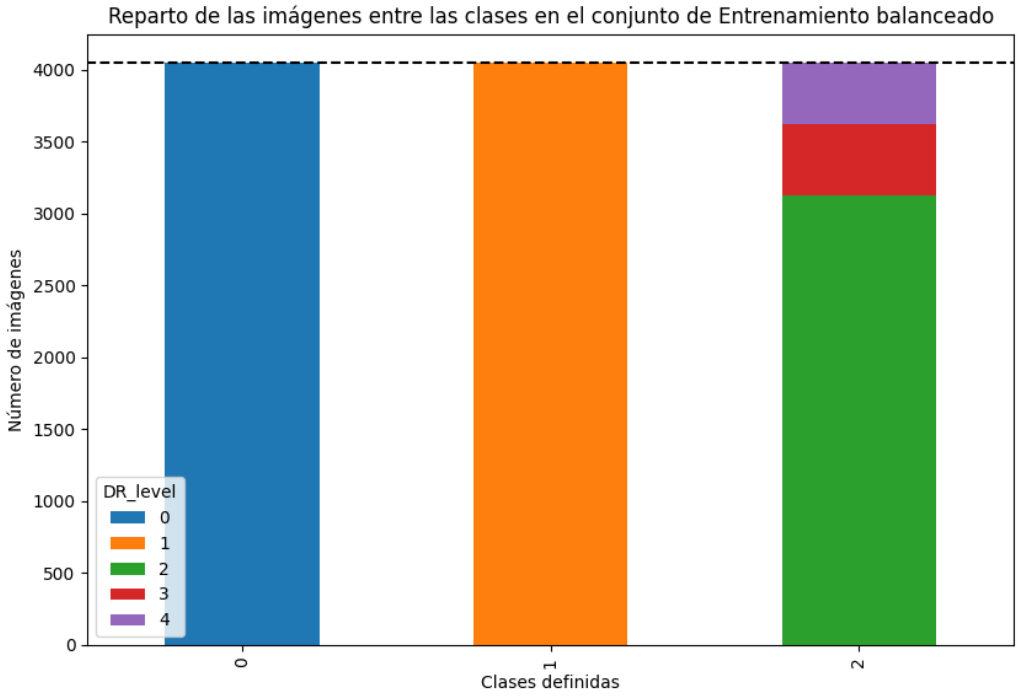


La imagen superior muestra la desbalanceada proporción entre las distintas clases, siendo muy notable el desequilibrio entre el número de imágenes de pacientes sanos y pacientes con los primeros indicios de la enfermedad (aproximadamente 10 imágenes de la primera por cada 1 de la segunda). Concretamente, 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 esta gran diferencia, en la etapa de entrenamiento sólo serán analizadas 4.049 imágenes de cada una de las clases. Esta cantidad se debe a que la clase minoritaria, compuesta únicamente por imágenes de RD leve, sólo posee esta cantidad de imágenes. De las clases mayoritarias, en cada época, se seleccionará una muestra aleatoria de 4.049 ejemplos, haciendo uso así de todas las imágenes. En total, en cada época, el *dataset* de entrenamiento se

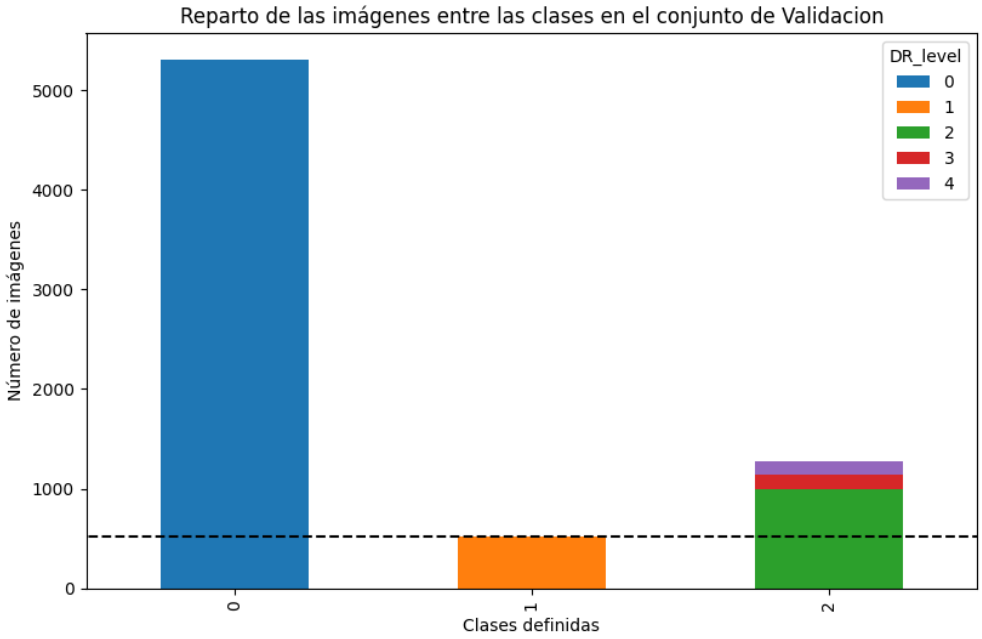
compondrá de 12.147 imágenes, siendo su distribución media durante todo el entrenamiento similar a esta:



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:

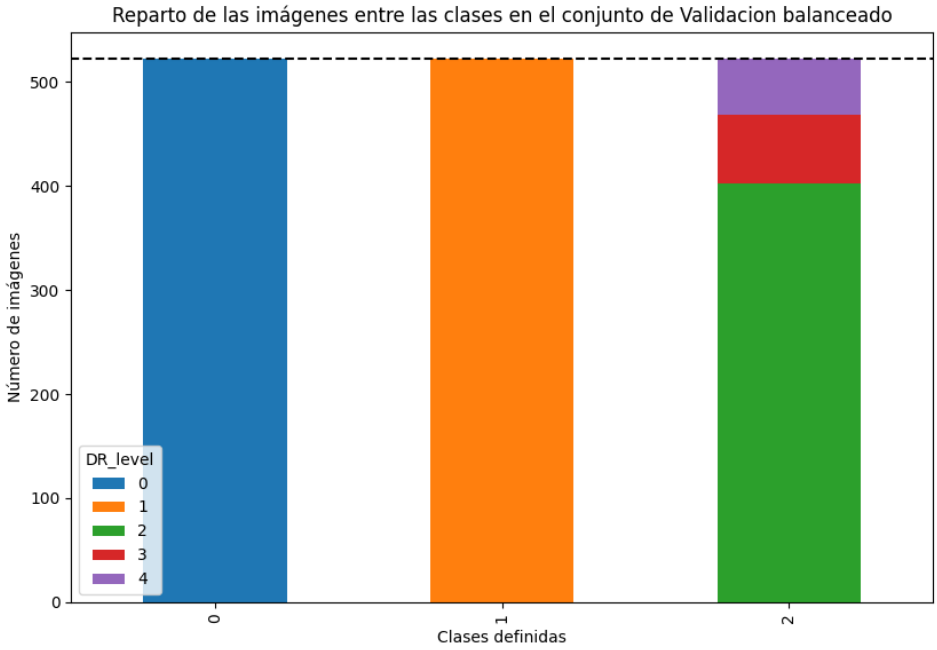
- Con nivel 0: 5.304 imágenes
- Con nivel 1: 522 imágenes
- Con nivel 2: 997 imágenes
- Con nivel 3: 138 imágenes
- Con nivel 4: 142 imágenes

Siendo su distribución la que se muestra a continuación:



Al igual que en el conjunto de entrenamiento, se tomarán exactamente, y de forma aleatoria (para mantener la distribución interna de niveles de DR en las clases definidas), el mismo número de imágenes de cada clase: 522 (número máximo de imágenes pertenecientes a la clase minoritaria), con la excepción de que, en este conjunto, las imágenes de las clases mayoritarias 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 se somete a una evaluación sobre un conjunto aparte de imágenes (el conjunto de validación).

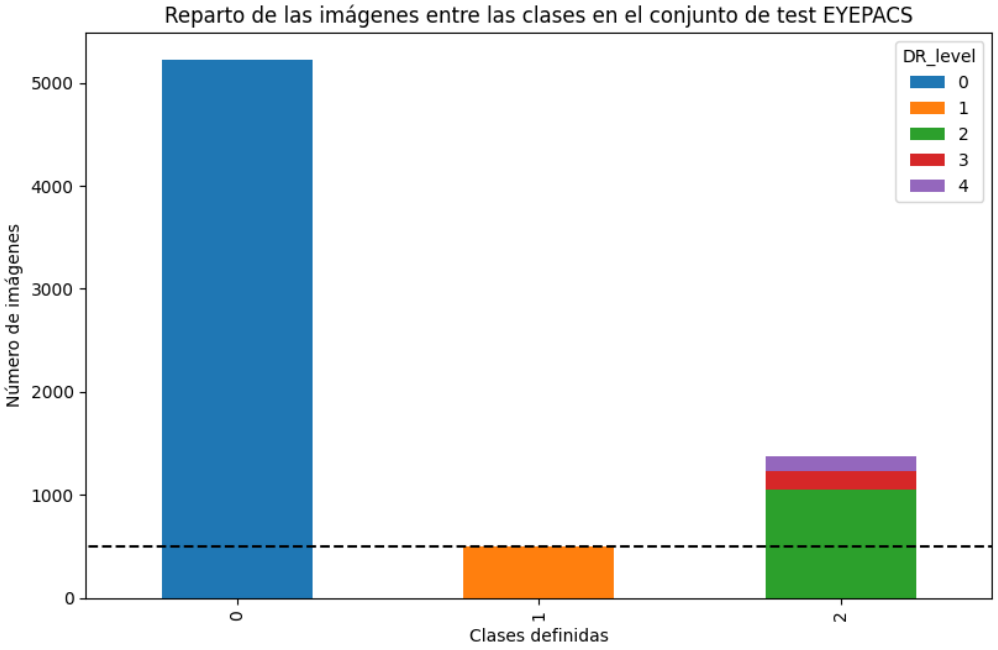
Mediante esta acción, se lleva a cabo un control más preciso del rendimiento real de la red. Por tanto, no es deseable que este conjunto pueda variar sus elementos conforme el entrenamiento avanza. El número total de imágenes de validación será de 1.566, que no se verán modificadas en ningún momento del entrenamiento. Nueva distribución tras el descarte de las imágenes queda de la siguiente forma:



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.

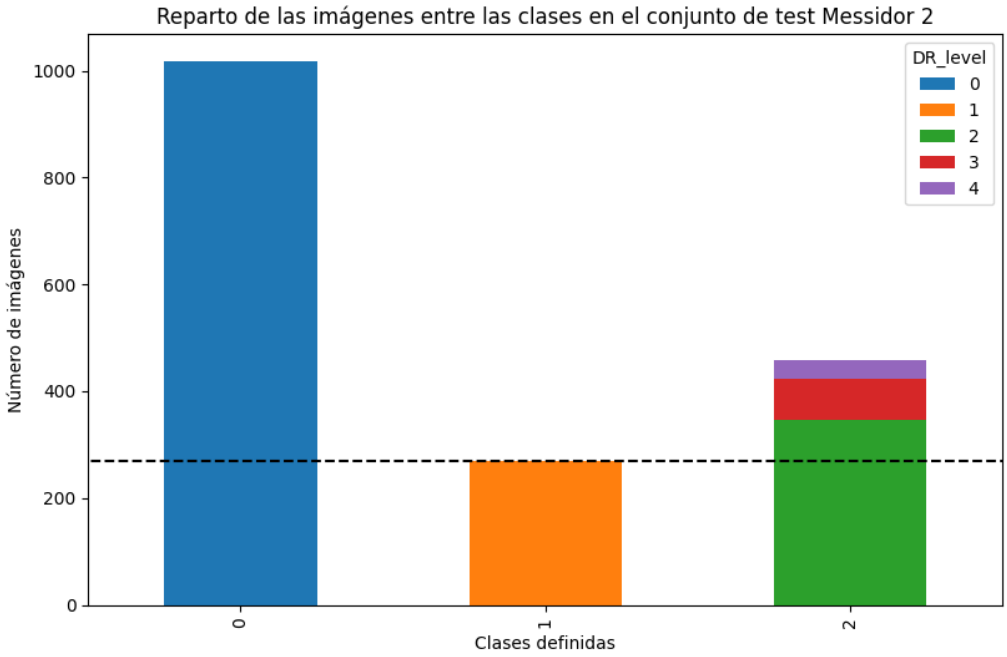
Esto hace que la distribución de las imágenes en el conjunto de test sea la siguiente:

- 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 diagnóstico 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

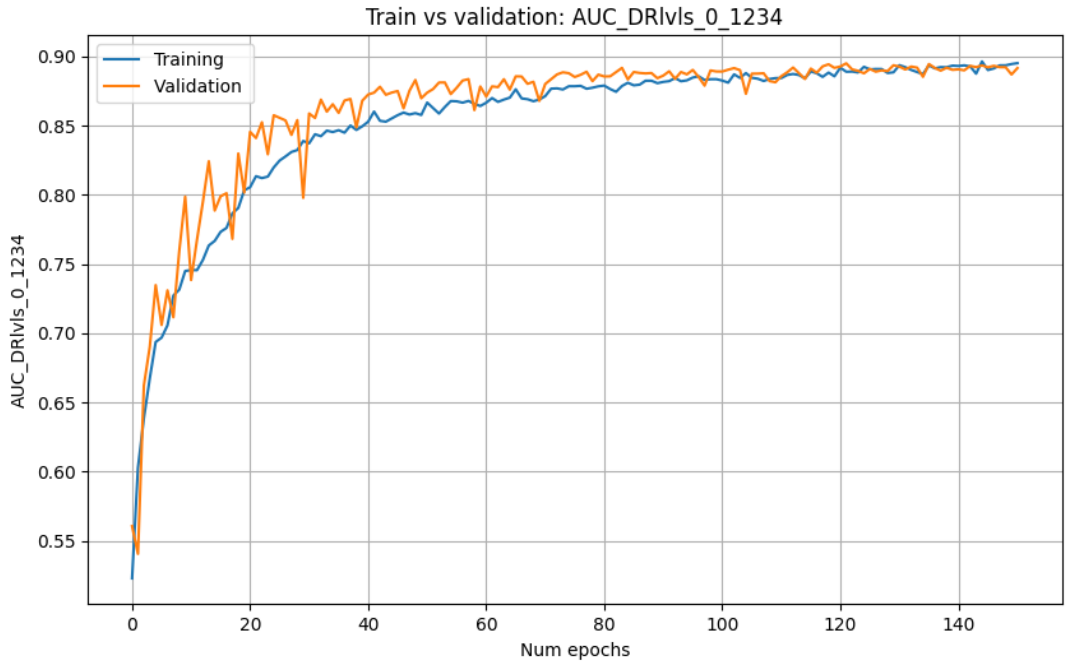


Nuevamente, la presencia de imágenes correspondientes a pacientes sanos es mayor que las demás clases, aunque en este conjunto, la diferencia no es tan marcada.

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 tomando como clase positiva que la imagen posea algún signo de enfermedad, es decir, leve o superior.

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 121 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

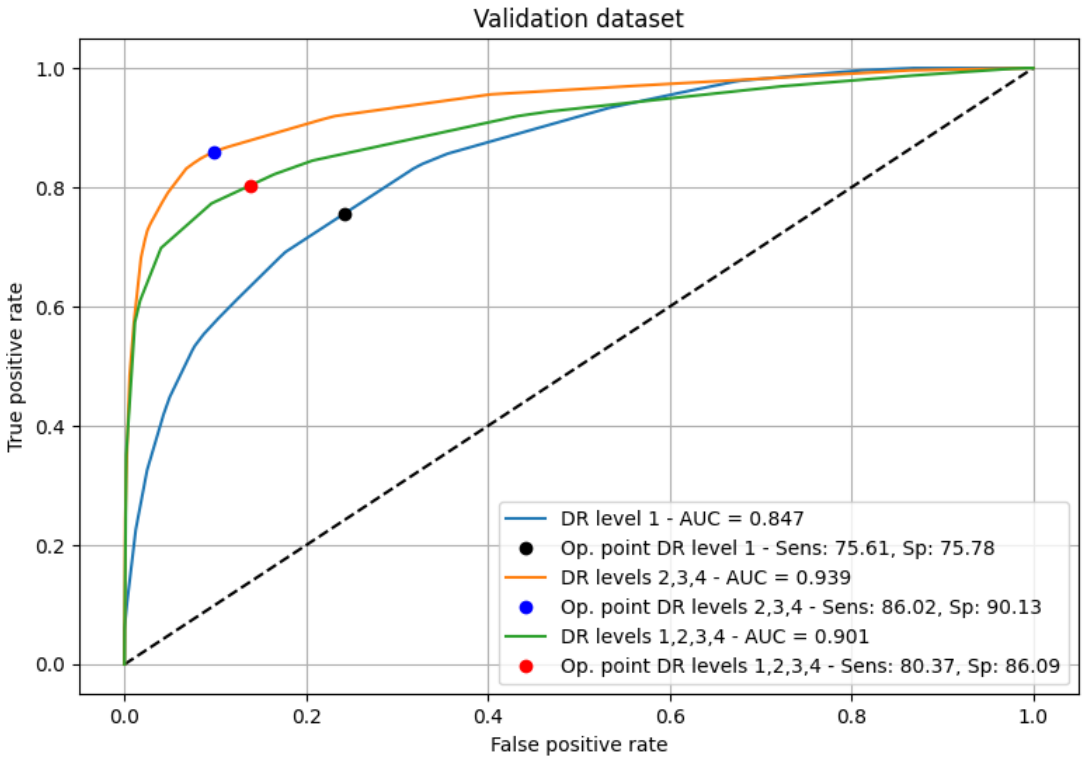
Como en este entrenamiento las imágenes fueron clasificadas en sanos, enfermedad leve y enfermedad moderada o peor, se puede medir el rendimiento de la red para la detección de todos los pacientes enfermos, de los pacientes con la enfermedad en un nivel más avanzado o sólo los pacientes que tengan los primeros indicios de RD. Esto da lugar a los siguientes resultados:

Detección de sólo los casos leves de RD (nivel 1). AUC = 0.8466					
Punto de operación	**Sensibilidad**	**Especificidad**	**Tasa de Falsos Negativos (TFN)**	**TFN Nivel 1**	**Umbrales**
98	98.85	25.00	1.15	1.15	0.036072
95	95.98	38.51	4.02	4.02	0.078156
93	93.30	46.74	6.70	6.70	0.108216
90	93.30	46.74	6.70	6.70	0.108216
85	85.63	64.46	14.37	14.37	0.200401
Punto más cercano	75.61	75.78	24.39	26.44	0.317065

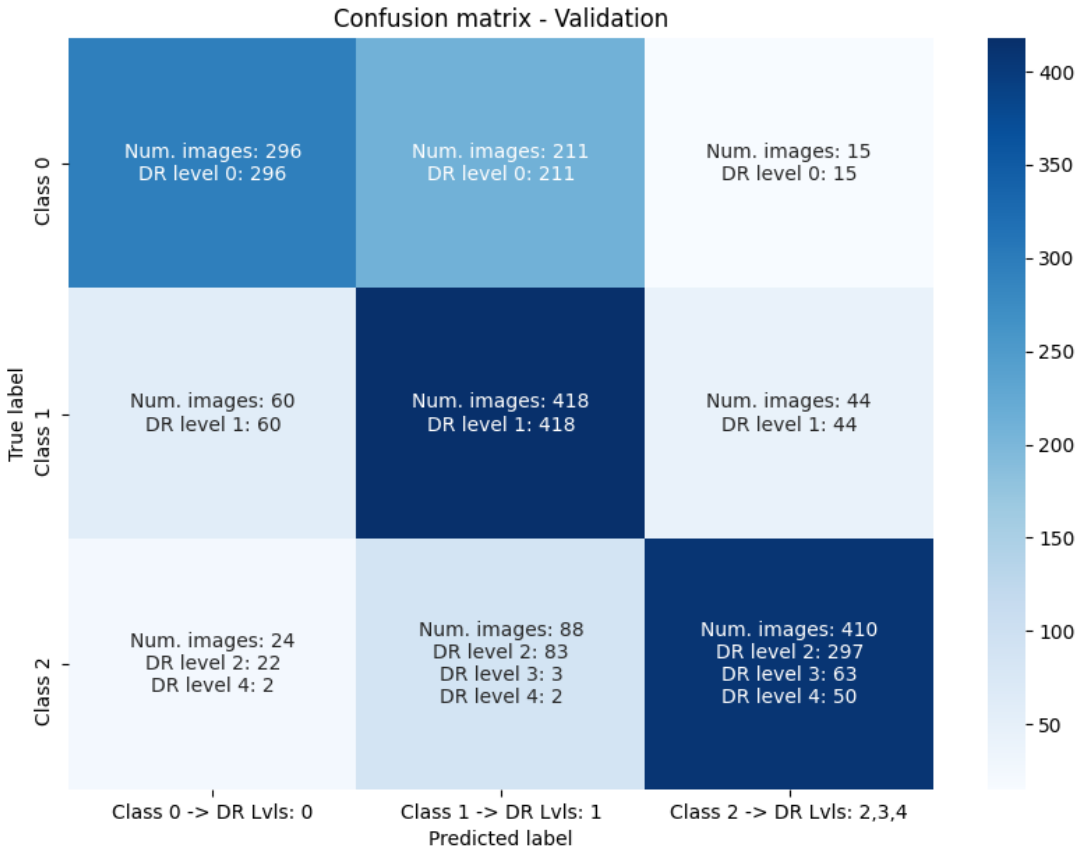
En el caso mostrado en la tabla superior, la información acerca de los falsos negativos no sería de gran utilidad debido a que, para obtener dichos resultados, se tuvo que considerar como clase positiva que la imagen padezca de RD en nivel 1, y como clase negativa, los niveles 2, 3 y 4, además de sanos.

Detección de sólo los casos moderados o peores (niveles 2, 3 y 4). AUC = 0.9387							
Punto de operación	**Sensibilidad**	**Especificidad**	**Tasa de Falsos Negativos (TFN)**	**TFN Nivel 2**	**TFN Nivel 3**	**TFN Nivel 4**	**Umbrales**
98	99.62	13.60	0.38	0.50	0.00	0.00	0.002004
95	95.59	59.67	4.41	5.22	0.00	3.70	0.030060
93	95.59	59.67	4.41	5.22	0.00	3.70	0.030060
90	91.38	78.26	8.62	9.95	3.03	5.56	0.090180
85	86.02	90.13	13.98	16.67	3.03	7.41	0.228457
Punto más cercano	86.02	90.13	13.98	16.67	3.03	7.41	0.228457

Detección de todos los casos de RD. AUC = 0.901								
Punto de operación	**Sensibilidad**	**Especificidad**	**Tasa de Falsos Negativos (TFN)**	**TFN Nivel 1**	**TFN Nivel 2**	**TFN Nivel 3**	**TFN Nivel 4**	**Umbrales**
98	98.75	13.22	1.25	1.15	1.74	0.00	0.00	0.066132
95	96.93	27.78	3.07	4.21	2.49	0.00	0.00	0.110220
93	96.93	27.78	3.07	4.21	2.49	0.00	0.00	0.110220
90	91.95	56.70	8.05	11.49	5.47	0.00	3.70	0.234469
85	91.95	56.70	8.05	11.49	5.47	0.00	3.70	0.234469
Punto más cercano	80.37	86.09	19.63	31.80	9.95	1.52	5.56	0.543256



En todas las tablas ha sido marcada la fila cuyo umbral garantizaba al menos un 90% de sensibilidad. Haciendo uso de estos umbrales, se ha procedido a la clasificación de todas las imágenes pertenecientes al conjunto de validación, asignando a cada imagen una de las tres clases existentes (sanos, enfermedad leve, enfermedad moderada o peor). Con esta clasificación, se ha construido la siguiente matriz de confusión:



Como se puede observar, el número de imágenes con patologías que han sido clasificadas como sanas han sido un total de:

- Para el nivel 1 de RD: 60 imágenes de las 522.
- Para el nivel 2: 22 imágenes de las 402.
- Para el nivel 3: ninguna imagen de las 66.
- Para el nivel 4: 2 imágenes de las 54.

El error de clasificación dado en casos de imágenes de niveles 2 o superior clasificadas como nivel 1 no tiene el mismo nivel de importancia ya que serían enviadas a revisión. De igual manera sucedería con aquellos sanos que se les ha catalogado como pacientes con la enfermedad, en una revisión futura serían descartados.

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

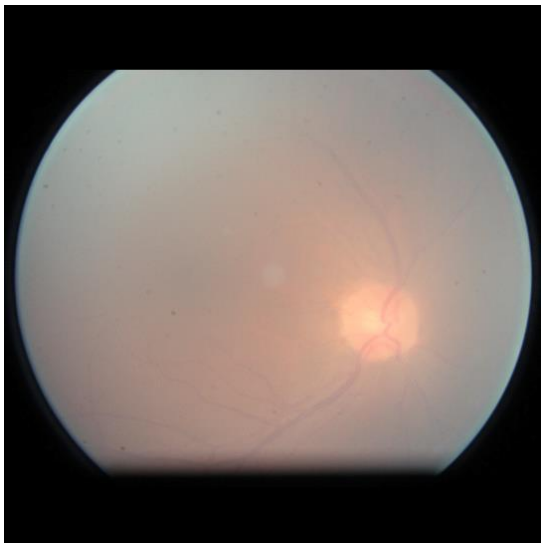
Nivel real de RD: 4

Nivel real de RD: 4

Nivel real de RD: 2

Nivel real de RD: 2

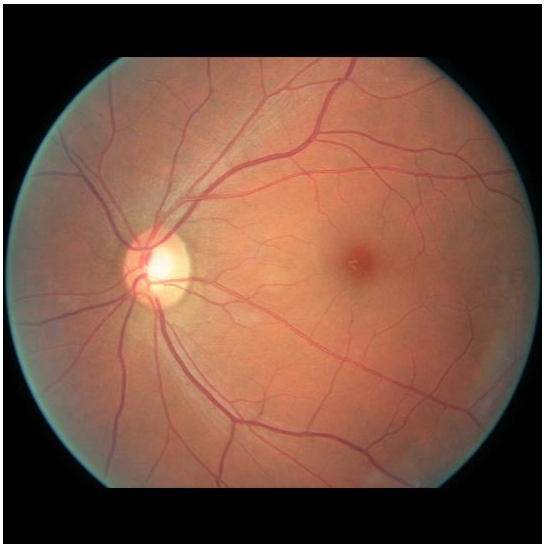
Nivel real de RD: 2



Nivel real de RD: 2

Nivel real de RD: 1

Nivel real de RD: 1



Trayendo los umbrales obtenidos en el conjunto de validación a los conjuntos de prueba, los resultados obtenidos son los siguientes. Nuevamente, se medirá la capacidad de la red de detectar sólo los pacientes con RD leve, sólo los pacientes con RD moderada o peor, y todos los pacientes con cualquier nivel de RD.

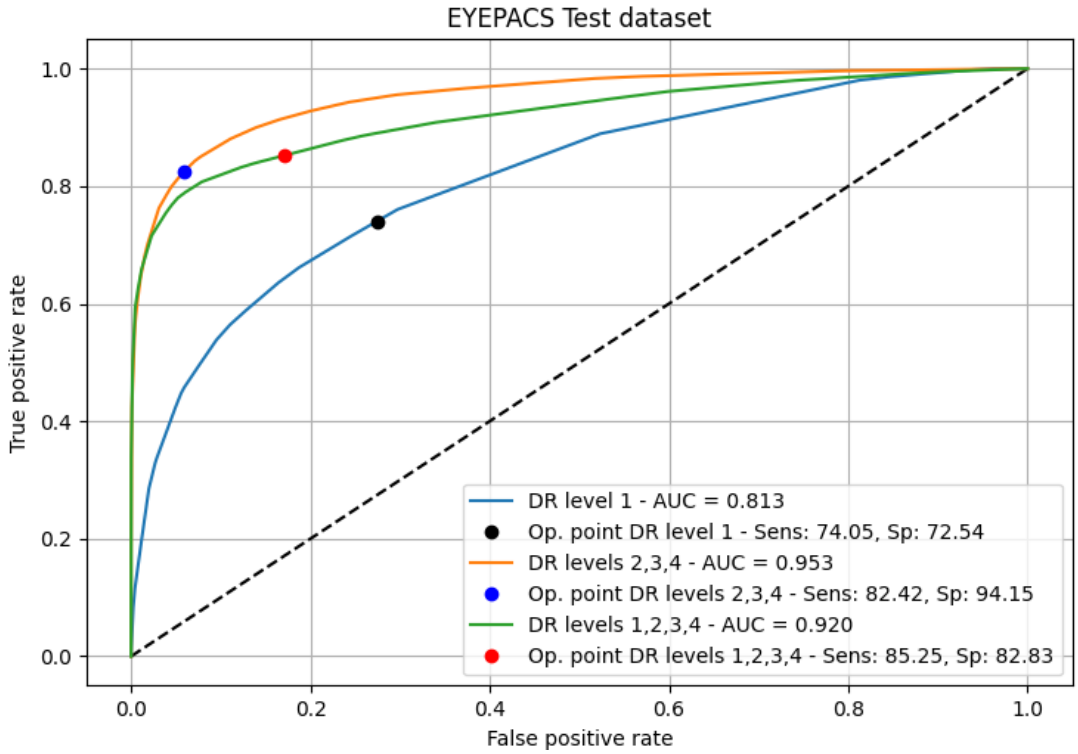
Resultados en el conjunto de testeo de EYEPACS

Detección de sólo los casos leves de RD (nivel 1). AUC = 0.8132					
Punto de operación	**Sensibilidad**	**Especificidad**	**Tasa de Falsos Negativos (TFN)**	**TFN Nivel 1**	**Umbrales**
98	98.85	13.54	1.15	1.39	0.036072
95	95.70	26.15	4.30	5.54	0.078156
93	93.02	34.66	6.98	7.72	0.108216
90	93.02	34.66	6.98	7.72	0.108216
85	84.62	55.23	15.38	17.82	0.200401
Punto más cercano	74.05	72.54	25.95	26.14	0.317065

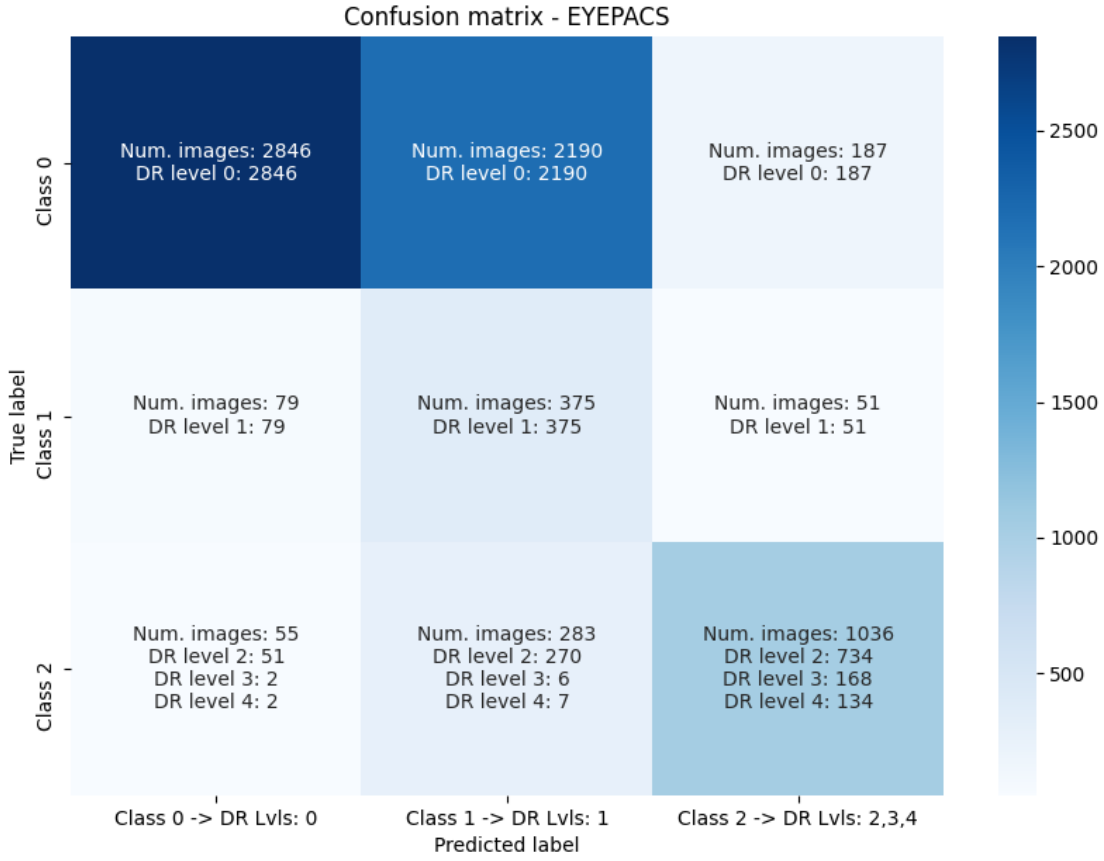
En el caso mostrado en la tabla superior, la información acerca de los falsos negativos no sería de gran utilidad debido a que, para obtener dichos resultados, se tuvo que considerar como clase positiva que la imagen padezca de RD en nivel 1, y como clase negativa, los niveles 2, 3 y 4, además de sanos. A destacar también que, como los umbrales utilizados son trasladados desde el conjunto de validación, es posible que no coincidieran con los umbrales que constituyen la ROC de este *dataset*. En ese caso, los valores de sensibilidad y especificidad fueron interpolados a través de los umbrales más cercanos, pero, en el caso de la tasa de falsos negativos por cada nivel de RD, éstos fueron calculados directamente sobre el conjunto. Por tanto, los resultados podrían no coincidir y ser ligeramente peores que la tasa global.

Detección de sólo los casos moderados o peores (niveles 2, 3 y 4). AUC = 0.9526							
Punto de operación	**Sensibilidad**	**Especificidad**	**Tasa de Falsos Negativos (TFN)**	**TFN Nivel 2**	**TFN Nivel 3**	**TFN Nivel 4**	**Umbrales**
98	99.64	20.11	0.36	0.47	0.00	0.00	0.002004
95	94.82	73.49	5.18	6.35	1.70	1.40	0.030060
93	94.82	73.49	5.18	6.35	1.70	1.40	0.030060
90	89.36	86.98	10.64	13.46	2.27	2.10	0.090180
85	82.42	94.15	17.58	21.99	2.84	4.20	0.228457
Punto más cercano	82.42	94.15	17.58	21.99	2.84	4.20	0.228457

Detección de todos los casos de RD. AUC = 0.9201								
Punto de operación	**Sensibilidad**	**Especificidad**	**Tasa de Falsos Negativos (TFN)**	**TFN Nivel 1**	**TFN Nivel 2**	**TFN Nivel 3**	**TFN Nivel 4**	**Umbrales**
98	99.09	14.16	0.91	2.57	0.57	0.00	0.70	0.066132
95	97.79	27.46	2.21	5.15	1.52	0.00	1.40	0.110220
93	97.79	27.46	2.21	5.15	1.52	0.00	1.40	0.110220
90	93.46	53.30	6.54	15.64	4.83	1.14	1.40	0.234469
85	93.46	53.30	6.54	15.64	4.83	1.14	1.40	0.234469
Punto más cercano	85.25	82.83	14.75	30.10	11.56	1.70	5.59	0.543256



Se han marcado las filas que contienen los umbrales que garantizaron, en el conjunto de validación, una sensibilidad de al menos un 90%. Con estos umbrales, se ha procedido a la clasificación de todas las imágenes pertenecientes al conjunto de validación, asignando a cada imagen una de las tres clases existentes (sanos, enfermedad leve, enfermedad moderada o peor). Con esta clasificación, se ha construido la siguiente matriz de confusión:



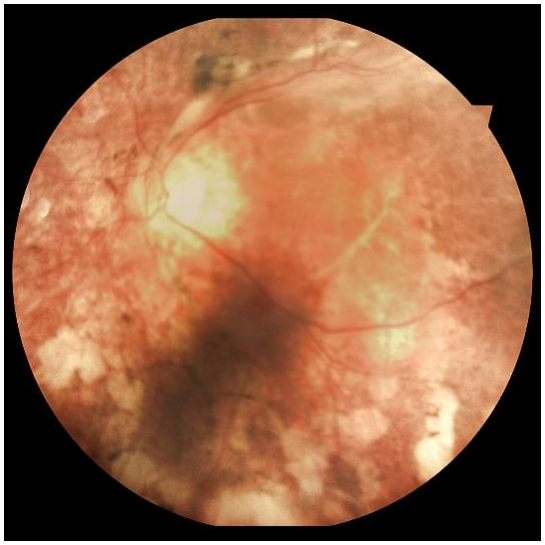
Como se puede observar, el número de imágenes con patologías que han sido clasificadas como sanas han sido un total de:

- Para el nivel 1 de RD: 79 imágenes de las 505.
- Para el nivel 2: 51 imágenes de las 1.055.
- Para el nivel 3: 2 imágenes de las 176.
- Para el nivel 4: 2 imágenes de las 143.

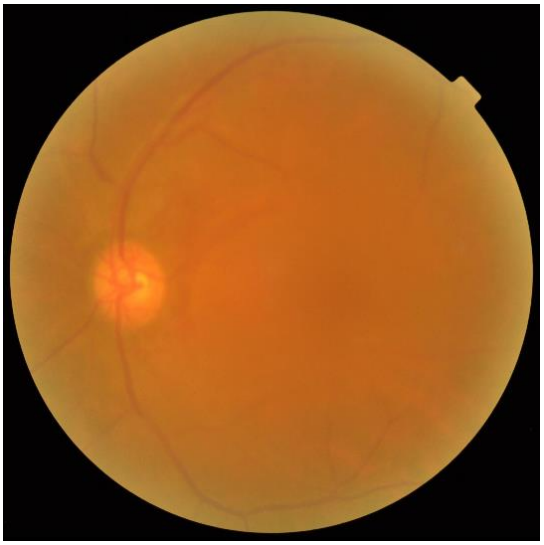
El error de clasificación dado en casos de imágenes de niveles 2 o superior clasificadas como nivel 1 no tiene el mismo nivel de importancia ya que serían enviadas a revisión. De igual manera sucedería con aquellos sanos que se les ha catalogado como pacientes con la enfermedad, en una revisión futura serían descartados.

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

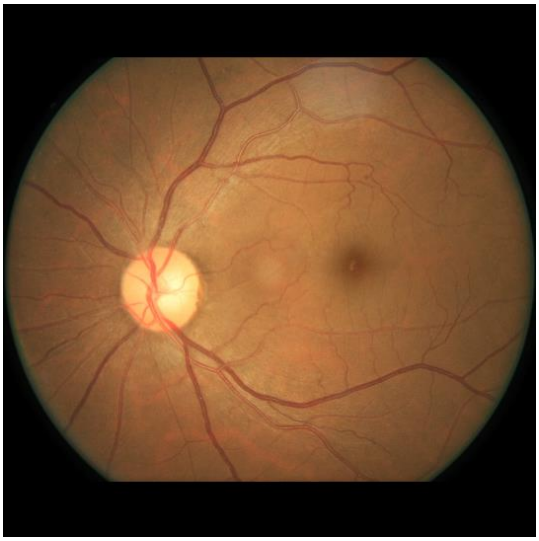
Nivel real de RD: 4



Nivel real de RD: 4

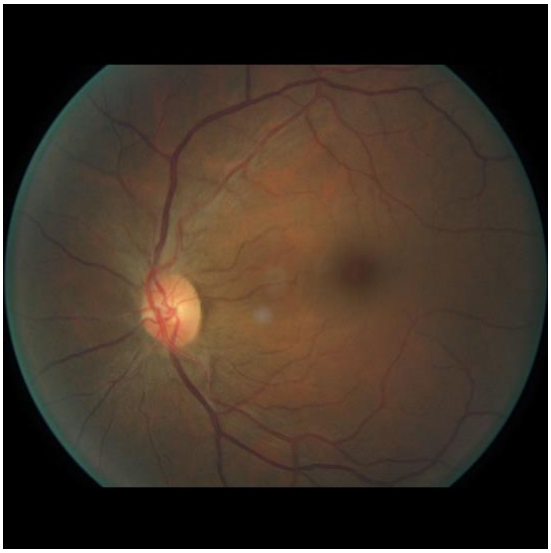


Nivel real de RD: 3



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

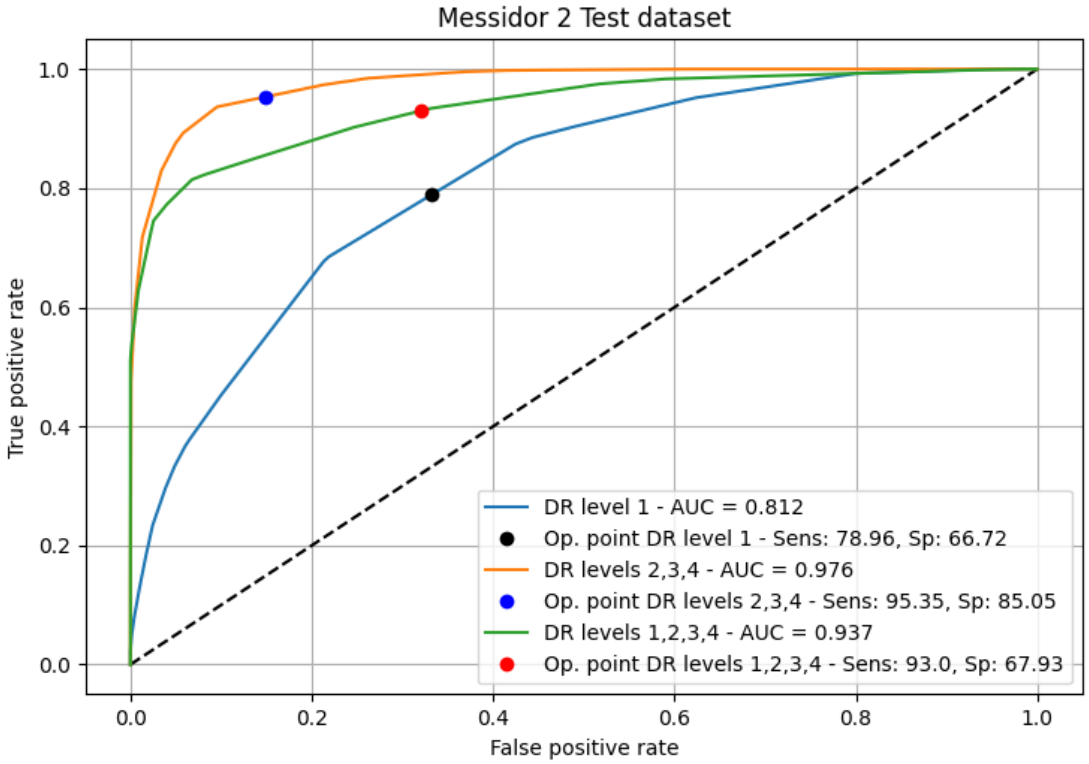
Resultados en el conjunto de testeo de Messidor-2

Detección de sólo los casos leves de RD (nivel 1). AUC = 0.812					
Punto de operación	**Sensibilidad**	**Especificidad**	**Tasa de Falsos Negativos (TFN)**	**TFN Nivel 1**	**Umbrales**
98	98.93	21.42	1.07	2.22	0.036072
95	95.51	36.24	4.49	4.81	0.078156
93	93.34	42.74	6.66	7.41	0.108216
90	93.34	42.74	6.66	7.41	0.108216
85	87.52	57.34	12.48	12.59	0.200401
Punto más cercano	78.96	66.72	21.04	22.22	0.317065

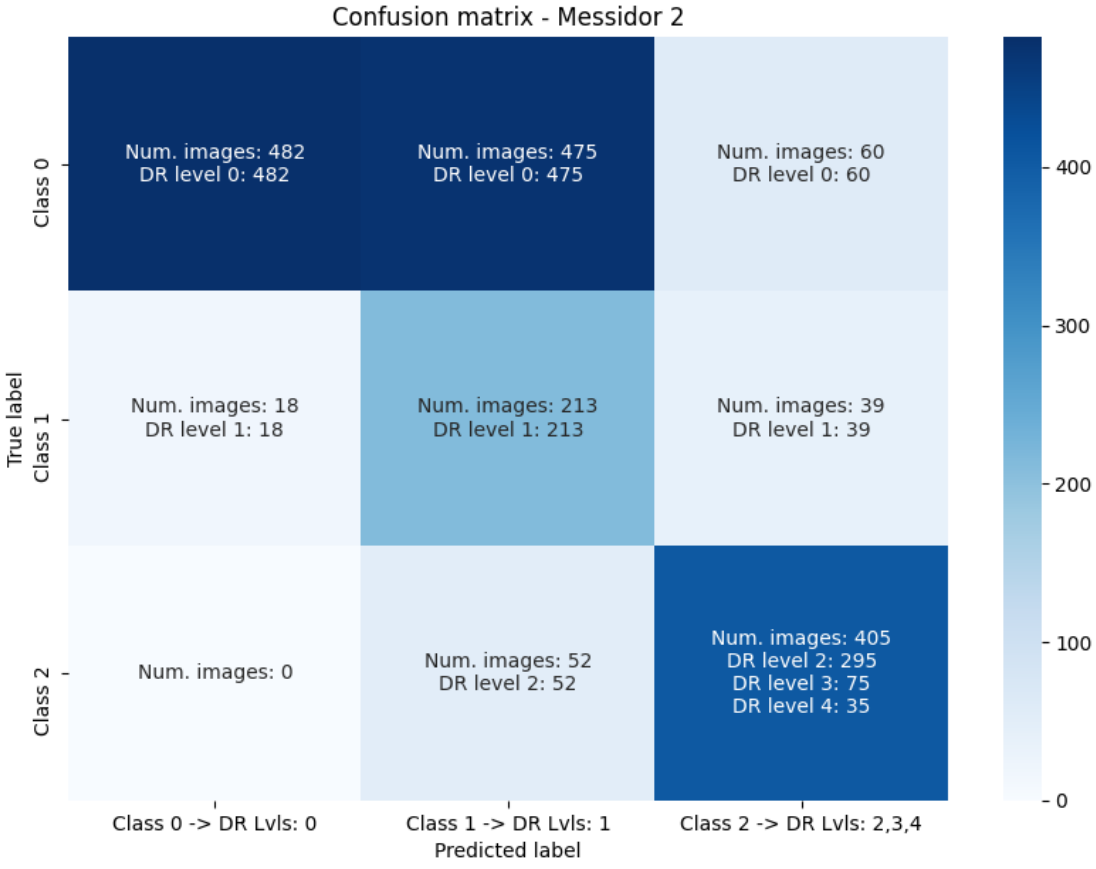
En el caso mostrado en la tabla superior, la información acerca de los falsos negativos no sería de gran utilidad debido a que, para obtener dichos resultados, se tuvo que considerar como clase positiva que la imagen padezca de RD en nivel 1, y como clase negativa, los niveles 2, 3 y 4, además de sanos. A destacar también que, como los umbrales utilizados son trasladados desde el conjunto de validación, es posible que no coincidieran con los umbrales que constituyen la ROC de este *dataset*. En ese caso, los valores de sensibilidad y especificidad fueron interpolados a través de los umbrales más cercanos, pero, en el caso de la tasa de falsos negativos por cada nivel de RD, éstos fueron calculados directamente sobre el conjunto. Por tanto, los resultados podrían no coincidir y ser ligeramente peores que la tasa global.

Detección de sólo los casos moderados o peores (niveles 2, 3 y 4). AUC = 0.9758							
Punto de operación	**Sensibilidad**	**Especificidad**	**Tasa de Falsos Negativos (TFN)**	**TFN Nivel 2**	**TFN Nivel 3**	**TFN Nivel 4**	**Umbrales**
98	100.00	5.48	0.00	0.00	0.00	0.00	0.002004
95	99.90	47.71	0.10	0.29	0.00	0.00	0.030060
93	99.90	47.71	0.10	0.29	0.00	0.00	0.030060
90	98.85	70.01	1.15	2.02	0.00	0.00	0.090180
85	95.35	85.05	4.65	7.20	0.00	0.00	0.228457
Punto más cercano	95.35	85.05	4.65	7.20	0.00	0.00	0.228457

Detección de todos los casos de RD. AUC = 0.9366								
Punto de operación	**Sensibilidad**	**Especificidad**	**Tasa de Falsos Negativos (TFN)**	**TFN Nivel 1**	**TFN Nivel 2**	**TFN Nivel 3**	**TFN Nivel 4**	**Umbrales**
98	99.48	15.02	0.52	2.59	0.00	0.00	0.00	0.066132
95	98.99	26.19	1.01	4.07	0.00	0.00	0.00	0.110220
93	98.99	26.19	1.01	4.07	0.00	0.00	0.00	0.110220
90	97.63	47.22	2.37	6.67	0.00	0.00	0.00	0.234469
85	97.63	47.22	2.37	6.67	0.00	0.00	0.00	0.234469
Punto más cercano	93.00	67.93	7.00	19.26	0.29	0.00	2.86	0.543256



Se han marcado las filas que contienen los umbrales que garantizaron, en el conjunto de validación, una sensibilidad de al menos un 90%. Con estos umbrales, se ha procedido a la clasificación de todas las imágenes pertenecientes al conjunto de validación, asignando a cada imagen una de las tres clases existentes (sanos, enfermedad leve, enfermedad moderada o peor). Con esta clasificación, se ha construido la siguiente matriz de confusión:



Como se puede observar, el número de imágenes con patologías que han sido clasificadas como sanas han sido un total de:

- Para el nivel 1 de RD: 18 imágenes de las 270.
- Para el nivel 2: ninguna imagen de las 347.
- Para el nivel 3: ninguna imagen de las 75.
- Para el nivel 4: ninguna imagen de las 35.

El error de clasificación dado en casos de imágenes de niveles 2 o superior clasificadas como nivel 1 no tiene el mismo nivel de importancia ya que serían enviadas a revisión. De igual manera sucedería con aquellos sanos que se les ha catalogado como pacientes con la enfermedad, en una revisión futura serían descartados.

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: 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.