

Etiquetado, clasificación y análisis de calidad de imagen para detección de retinopatía diabética usando modelos convolucionales profundos

Pedro de J. Bermejo-Guerrero¹, Abraham Sánchez²,
E. Ulises Moya-Sánchez^{1,2}, Ulises Cortés³

¹ Universidad Autónoma de Guadalajara,
Posgrado en Ciencias Computacionales,
México

² Gobierno de Jalisco,
Coordinación General de Innovación Gubernamental,
México

³ Barcelona Supercomputing Center,
España

eduardo.moya@jalisco.gob.mx

Resumen. La retinopatía diabética (RD) es la principal causa de ceguera en el mundo en edad laboral. La detección-tratamiento temprano ha demostrado ser efectiva para disminuir las complicaciones visuales (y la ceguera) de los pacientes diabéticos. Los modelos profundos han mostrado recientemente que pueden ayudar con la detección temprana analizando masivamente un gran número de imágenes. Uno de los problemas con esos modelos en su aplicación en el mundo real es que la calidad de la imagen puede afectar significativamente el desempeño de los modelos haciendo inviable la evaluación automática. En este trabajo presentamos tres elementos principales i) Etiquetado de calidad: criterios, herramientas, etiquetas en imágenes públicas; ii) comparación de modelos convolucionales y iii) análisis sobre el desempeño de los modelos con degradaciones sintéticas. Los resultados muestran que se puede alcanzar una exactitud del 97 % y los análisis ayudan a entender los tipos de degradaciones que afectan más a las redes. Estos modelos y resultados serán especialmente útiles para quienes tengan interés de desplegar modelos de RD en ambientes reales.

Palabras clave: Retinopatía diabética, calidad de imagen, modelos convolucionales profundos.

Labeling, Classification and Image Quality Analysis for Detection of Diabetic Retinopathy Using Deep ConvNets Models

Abstract. Diabetic retinopathy (DR) is the leading cause of blindness in the working-age population worldwide. Early detection and treatment have been

shown to be effective in reducing visual complications (and blindness) in diabetic patients. Deep learning models have recently demonstrated that they can aid in early detection by analyzing a massive number of images. One problem with these models in real-world applications is that image quality can significantly impact model performance, making automatic evaluation unfeasible. In this work, we present three main elements: i) Quality labeling - criteria, tools, and labels in public images; ii) Comparison of convolutional models; and iii) Analysis of model performance with synthetic degradations. The results show that an accuracy of 97% can be achieved, and the analyses help understand the types of degradation that most affect the networks. These models and results will be especially useful for those interested in deploying DR models in real-world environments.

Keywords: Diabetic retinopathy, deep Conv-Nets, image quality.

1. Introducción

En los últimos años, el aprendizaje profundo (AP) o Deep Learning se ha convertido en una poderosa herramienta para clasificar imágenes médicas [14]. En Retinopatía Diabética (RD) se han usado exitosamente las redes neuronales convolucionales (ConvNets) para la clasificación del nivel de RD en las imágenes de fondo de ojo [7, 15, 8]. Sin embargo, el desempeño de estos modelos puede disminuir significativamente cuando se usan datos del mundo real [19, 1, 8, 12, 11].

Aunque los sistemas de AP generalmente se entrenan y prueban en conjuntos de datos de imágenes de alta calidad, no se puede asumir que las imágenes del mundo real sean de calidad aceptable. Además, se ha demostrado que los modelos profundos convolucionales son muy sensibles a cambios de calidad de imagen, por ejemplo al cambio de: contraste, borrosidad y ruido [6]. Por lo tanto, es necesario tener herramientas para una evaluación de la calidad de la imagen [15].

En este trabajo presentamos tres elementos principales para poder evaluar y clasificar la calidad de imágenes en el contexto específico de la RD: i) etiquetado de imágenes públicas: herramientas, criterios específicos para la RD; ii) comparación de modelos convolucionales y iii) análisis sobre el desempeño de los modelos con degradaciones sintéticas.

Una de las contribuciones de este trabajo es ir más allá de los criterios clásicos de calidad de imagen: ruido, contraste, definición, agregando la presencia de elementos anatómicos: nervio óptico, y macula los cuales son fundamentales para detectar el nivel de RD de manera más confiable (incluso para el médico especialista).

Además, consideramos que es una contribución presentar una comparación del desempeño de varios modelos convolucionales. Por último, nuestro análisis con imágenes degradadas sintéticamente nos ayuda a evaluar el desempeño de los clasificadores en ambientes controlados, algo que no hemos visto en otros trabajos que abordan este tema.

Algunos trabajos similares son por ejemplo Zago ([20]) y Chalakkal ([3]), quienes obtienen un desempeño inferior del 98 % y 91.4 % en exactitud respectivamente. Sin embargo no toman en cuenta análisis de desempeño después del entrenamiento.

Además por desgracia no podemos comparar directamente estos trabajos ya que se usan datos diferentes y los códigos no están disponibles. Actualmente desarrollamos un conjunto de modelos para datos de México y las pruebas preliminares (sólo con datos públicos) muestran buenos resultados. Esperamos que estos resultados ayuden a evitar datos fuera de la distribución (por ejemplo datos de mala calidad) y así tener modelos más confiables para su aplicación en el mundo real.

2. Conceptos y fundamentos

Para etiquetar la calidad de imagen en RD no sólo usamos conceptos básicos de calidad de imagen, como ruido, contraste, nitidez, sino que además agregamos aspectos particulares RD como son artefactos o lesiones visibles y aspectos anatómicos particulares únicos de estas imágenes. A continuación enlistamos los conceptos de calidad de imagen más relevantes para el etiquetado de imagen de fondo de ojo con RD:

1. El *ruido* (cociente señal a ruido) es una variación aleatoria de los valores de intensidad de una imagen. La explicación del fenómeno tiene su origen frecuentemente en la poca señal detectada por el sensor (por ejemplo poca luz) [2].
2. El *rango dinámico*, es el intervalo de niveles de luz que el sensor (cámara) puede capturar [9]. En general los rango dinámicos de las imágenes médicas son mayores (> 8 bits) que las de las imágenes convencionales. Además, el rango dinámico está ligado con *el contraste*, que está vinculada a la diferencia relativa entre los valores intensidad de imagen en dos puntos de la imagen.
3. La *nitidez* determina el nivel de detalle (objeto más pequeño que se puede separar) en una imagen.
4. Los *artefactos* en RD son objetos, texturas o regiones de saturación que no corresponden al objeto real. Por ejemplo, elementos inherentes de la cámara, o destellos de luz en las imágenes de fondo de ojo.
5. Aspectos del *campo de visión* (anatómicos o lesiones en RD) tales como el nervio óptico, macula y lesiones como microaneurismas. Según la experiencia de médicos especialistas estos dos elementos son fundamentales y si no se encuentran debería de repetirse la captura o evaluarse para otras enfermedades.

3. Datos

3.1. Etiquetar imágenes

Se etiquetaron imágenes del conjunto de datos EyePACS-Kaggle [4], IDRID [13] y MESSIDOR [5] por tres personas. Usamos varios conjuntos de datos independientes y varias personas para disminuir la probabilidad de sesgos en el etiquetado y en los datos.

3.2. Clasificación

En la Tabla 1 mostramos como se dividieron el número de imágenes en los conjuntos de entrenamiento, validación y pruebas.

Tabla 1. División de conjunto de datos.

Conjunto	Buena calidad	Mala calidad
Entrenamiento	715	685
Validación	146	125
Pruebas	136	167
Total	997	977

3.3. Análisis de desempeño

Para analizar que tan sensibles eran estos modelos, se usaron 1,813 imágenes de buena calidad que no estaban en el conjunto de entrenamiento, validación o pruebas. Se eligieron tres tipo de degradaciones de imagen: blur (desenfoco), gauss noise (ruido) y random fog (artefactos y cambios de contraste) mediante el uso de Alumentations⁴.

4. Metodología y diseño experimental

4.1. Etiquetado de imágenes

Se tomaron en cuenta cinco factores para hacer el etiquetado de calidad de las imágenes de fondo de ojo de retinopatía (ver Figura 1):

1. Artefactos visibles. Se encuentran o no artefactos en el campo visual.
2. Nitidez. Distinción (binario) de los elementos anatómicos (separables). Por ejemplo, se pueden distinguir microaneurismas, neovasos y capilares hemorragias, exudados.
3. Campo de visión. Detección de disco óptico y mácula.
4. Lesiones graves. Existencia de problemas graves como hemorragias, dilatación venosa, desprendimientos, cirugías. En nuestra experiencia era importante etiquetar imágenes con estas lesiones ya que algunas de estas lesiones se puede confundir con artefactos.
5. Evaluación general. Clasificación evaluable (imagen de calidad), no evaluable (imagen de mala calidad).

El proceso de etiquetado fue realizado por tres personas y la etiqueta final fue dada por la mayoría. En la figura 1 se presenta un ejemplo de la herramienta interna utilizada para etiquetar las imágenes de fondo de ojo. Es importante notar que los factores antes mencionados se presentan como botones binarios que ayudan al usuario a etiquetar más rápidamente los factores. También queremos destacar que el tamaño de la imagen (número de píxeles) fue el tamaño original, ya que nos permitía ver con más detalle las lesiones o artefactos pequeños dentro de la imagen.

4.2. Clasificación de calidad

Se eligieron cuatro redes para comparar su rendimiento: InceptionV3 [18], MobilenetV2 [16], Resnet50 [10] y VGG19 [17]. Se eligieron estas redes ya que han tenido buenos resultados en la clasificación de imágenes para tareas similares.

⁴ <https://alumentations.ai/>

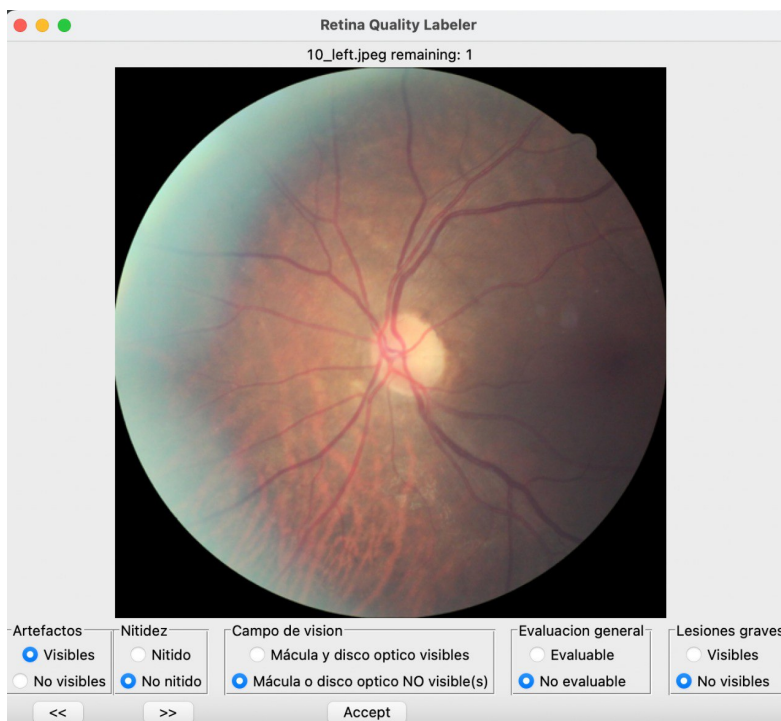


Fig. 1. Herramienta que se usó para etiquetar imágenes de fondo de ojo.

Sin embargo su tamaño (número de parámetros) y arquitecturas diferentes nos motivó a comparar su desempeño. A continuación se describe una lista de las arquitecturas y sus características más importantes.

- InceptionV3: 159 capas y 23.8 millones de parámetros.
- MobilenetV2: 53 capas y 3.4 millones de parámetros.
- Resnet50: 50 capas y 25.6 millones de parámetros.
- Vgg19: 19 capas y 143 millones de parámetros.

Podemos ver ciertas diferencias, InceptionV3 tiene un gran numero de capas pero un número medio de parámetros, MobilenetV2 tiene pocas capas y pocos parámetros siendo la red más ligera, Resnet50 tiene pocas capas y un número medio de parámetros, por último Vgg19 tiene muy pocas capas pero una gran cantidad de parámetros, siendo la red más pesada.

Los entrenamiento fueron usando los modelos pre-entrenados (imagenet) por lo que fue necesario sólo entrenar durante 20 épocas. Las imágenes de entrada tienen un resize de width=300, height=300 y una normalización con valores mean=[0, 0, 0], std=[1, 1, 1]. No se realizó ningún tipo de aumento de datos, ya que en estas imágenes las modificaciones más simples pueden alterar su clasificación; una imagen de buena calidad podría cambiar a ser una de mala calidad.

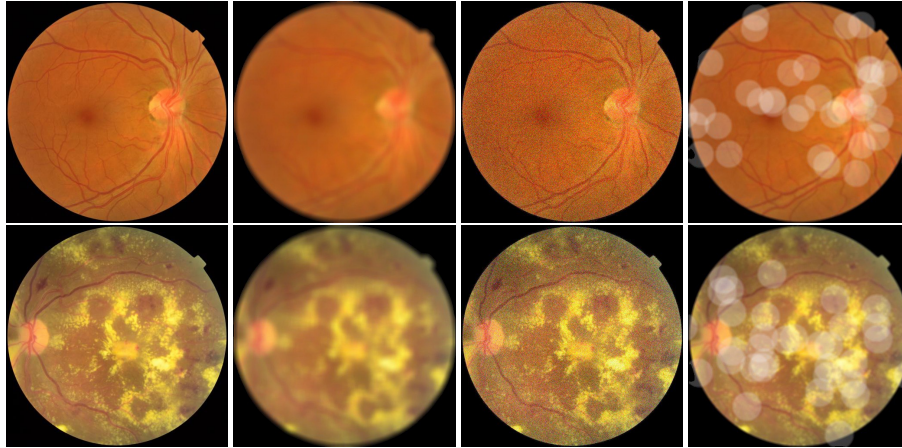


Fig.2. De izquierda a derecha se muestra: la imagen original, con borrosidad, ruido y niebla aleatoria.

Además se modificó la última capa de acuerdo al número de clases (2). Para entrenar se usó Pytorch como framework de AP. Para evaluar el desempeño de las redes convolucionales de clasificación de calidad de imagen, se usaron las siguientes métricas:

$$Exactitud = \frac{TP + TN}{TP + TN + FP + FN}, \quad (1)$$

$$Sensibilidad = \frac{TP}{TP + FN}, \quad (2)$$

$$Especificidad = \frac{TN}{TN + FP}, \quad (3)$$

En que TP = True Positive, TN = True Negative, FP = False Negative, y FN = False Positive.

4.3. Análisis de desempeño

El análisis de desempeño nos ayuda a evaluar la robustez/sensibilidad de los modelos ante tres ataques o degradaciones de imágenes. Se compara el intervalo de confianza con las imágenes originales (good) de buena calidad y se compara la distribución cuando se hace la degradación sintética de las imágenes. En la Figura 2 se muestran dos imágenes y sus degradaciones sintéticas, borrosidad (blur), ruido (noise), y niebla aleatoria (random fog).

5. Resultados y análisis

5.1. Etiquetado de imágenes

El principal resultado de esta subsección es haber creado una herramienta simple para etiquetar offline las imágenes de fondo de ojo con los criterios específicos de la RD.

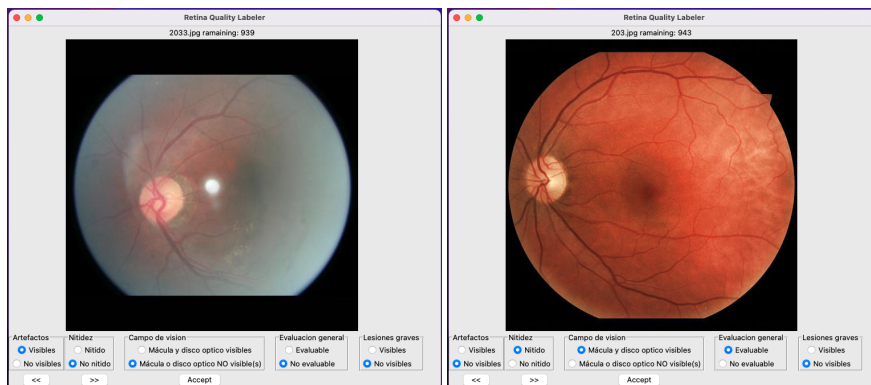


Fig. 3. Ejemplo de la herramienta que se usó para etiquetar imágenes.

En la Figura 3 se muestran dos ejemplos del etiquetado de imágenes. Los ejemplos son imágenes de buena calidad y con artefactos. El segundo resultado más importante de esta subsección es que se etiquetaron más de tres mil imágenes por tres personas tomando en cuenta aspectos cualitativos de la imagen.

Esperamos que esto reduzca los sesgos en las etiquetas de calidad y ayuden a un mejor aprendizaje del modelo. El aprendizaje que tuvimos al etiquetar estas imágenes nos ayudó a definir parámetros reales (acertados) para degradar las imágenes sintéticas (ver subsección Análisis de desempeño).

Es importante mencionar que las etiquetas estarán disponible públicamente en nuestro repositorio⁵. Esperamos que estas herramientas y etiquetas sirvan para que otras personas o grupos de investigación y además puedan comparar nuestro trabajo e incluso puedan pre-etiquetar sus imágenes de fondo de ojo.

5.2. Clasificación de calidad

En la Tabla 2 se muestran los resultados del entrenamiento de cuatro redes. de acuerdo a los resultados de la tabla, la red que tiene mejor exactitud, especificidad (Espe) y sensibilidad (Sen) es la Vgg19. Sin embargo los tiempos de entrenamiento y predicción son los más altos de todos los modelos.

Estos modelos como mencionamos anteriormente se han usado para tareas similares con gran éxito. Para poder entender porqué este modelo Vgg19 funciona mejor proponemos el análisis que hacemos en la siguiente sección. Además, para la explicabilidad, debemos realizar pruebas donde se muestra el gradiente de las últimas capas con imágenes de otro conjunto de datos.

5.3. Análisis de desempeño

En la Tabla 3 se presenta el porcentaje de imágenes clasificadas como mala calidad después de una degradación. El mayor porcentaje se asocia con una red más sensible a una degradación.

⁵ <https://github.com/PedroBermejo/retinopathy>

Tabla 2. Comparación de desempeño de clasificación de calidad de imagen de fondo de ojo. En que T. Entr es el tiempo de entrenamiento, T. Pred es el tiempo de predicción, Sen es la *Sensibilidad* y Espe es la *Especificidad*.

Modelo	Exactitud	Sen	Espe	T. Entr. (s)	T. Pred (s)
InceptionV3 [18]	0.91	0.91	0.91	605	52
MobileNetV2[16]	0.89	0.90	0.89	382	32
ResNet50 [10]	0.95	0.95	0.95	556	68
Vgg19[17]	0.97	0.98	0.95	983	145

Tabla 3. Porcentaje de imágenes que se clasificaron como de mala calidad después de las degradaciones. Un porcentaje mayor representa un mayor cambio o sensibilidad ante esas degradaciones.

Modelo	Borrosidad	Ruido	Niebla aleatoria
InceptionV3 [18]	62 %	4 %	11 %
MobilenetV2 [16]	47 %	4 %	11 %
Resnet50 [10]	9 %	1 %	6 %
Vgg19 [17]	95 %	8 %	76 %

Se puede observar que la borrosidad (blur) y niebla aleatoria (random fog) son las degradaciones que más afectan a las redes. Por otra parte el ruido es la degradación que menos afecta a las redes. Se puede observar que la red Vgg19 es la red más sensible a cada uno de los tipos de degradación. De acuerdo a este análisis y a los experimentos de clasificación consideramos que Vgg19 es red más sensible para esta tarea a pesar de ser la red más pesada y con mayor tiempo de entrenamiento y predicción.

En la Figura 4 se presenta la comparación de la distribución del intervalo de confianza (IC) mediante cajas y bigotes. El IC = 0 representa imágenes clasificadas con buena calidad mientras que el valor IC = 1 es el correspondiente a las clasificadas de mala calidad. Se puede observar como las imágenes originales (etiquetada como *referencia*) su distribución es cercana al cero.

Después de degradar esas imágenes podemos ver que el intervalo de confianza se distribuye con valores más cercanos a uno en el caso de Vgg10 e InceptionV3. Esta figura se puede considerar como representación gráfica de la tabla 3 y nos indica que tan sensibles son las redes a los tipos de degradación.

6. Conclusiones

En este trabajo presentamos tres elementos principales: etiquetado, clasificación y análisis. El etiquetado de calidad de imagen presenta los elementos y herramientas en el contexto específico de RD, ya que consideramos en el proceso de etiquetado los elementos de calidad de imagen y elementos anatómicos asociados al fondo de ojo y lesiones vinculadas a la enfermedad de RD.

Además, presentamos la comparación del desempeño de clasificación cuatro modelos convolucionales.

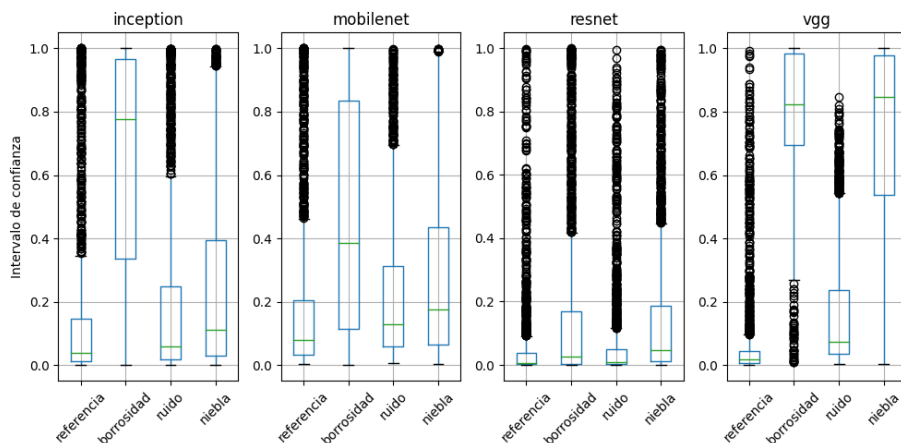


Fig. 4. Análisis del desempeño de clasificación con imágenes sintéticas. Se presenta en intervalo de confianza de las imágenes de referencia y de las mismas imágenes degradadas sintéticamente. De izquierda a derecha referencia (good quality), borrosidad (blur), ruido (gauss noise), niebla (random fog).

El modelo Vgg19 presenta mejores métricas de exactitud, especificidad y sensibilidad, pero con el mayor tiempo de entrenamiento e inferencia. Por último presentamos un análisis de los modelos cuando se hace la inferencia con imágenes degradadas sintéticamente.

El análisis del desempeño nos confirma que el modelo Vgg19 es el más sensible ante cambios de borrosidad o niebla aleatoria. Un modelo con más sensibilidad nos ayuda a una mejor clasificación de imágenes de buena o mala calidad.

En contraste un modelo con muy poca sensibilidad como lo es Resnet50 podría ser muy buena opción para la etapa de clasificación de niveles de retinopatía.

Estos modelos y resultados serán especialmente útiles para quienes tengan interés de desplegar modelos en ambientes reales ya que consideran aspectos de robustez y rapidez (menor tiempo de inferencia).

Consideramos que en un futuro este análisis nos servirá para implementar un sistema de modelos profundos con imágenes regionales que ya han sido colectadas por el Gobierno del Estado de Jalisco.

Referencias

1. Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., Vardoulakis, L. M.: A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–12 (2020) doi: 10.1145/3313831.3376718
2. Bushberg, J. T., Seibert, J. A., Leidholdt, E. M., Boone, J. M.: The essential physics of medical imaging, Lippincott Williams and Wilkins (2011)

3. Chalakkal, R. J., Abdulla, W. H., Thulaseedharan, S. S.: Quality and content analysis of fundus images using deep learning. *Computers in Biology and Medicine*, vol. 108, pp. 317–331 (2019) doi: 10.1016/j.combiomed.2019.03.019
4. Cuadros, J., Bresnick, G.: EyePACS: An adaptable telemedicine system for diabetic retinopathy screening. *Journal of diabetes science and technology*, vol. 3, no. 3, pp. 509–516 (2009) doi: 10.1177/193229680900300315
5. Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., Gain, P., Ordonez, R., Massin, P., Erginay, A., Charton, B., Klein, J. C.: Feedback on a publicly distributed database: The Messidor database. *Image Analysis & Stereology*, vol. 33, no. 3, pp. 231–234 (2014) doi: 10.5566/ias.1155
6. Dodge, S., Karam, L.: Understanding how image quality affects deep neural networks. In: 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), pp. 1–6 (2016) doi: 10.1109/QoMEX.2016.7498955
7. Fan, R., Liu, Y., Zhang, R.: Multi-scale feature fusion with adaptive weighting for diabetic retinopathy severity classification. *Electronics*, vol. 10, no. 12 (2021) doi: 10.3390/electronics10121369
8. Gaspar-González, B., Sánchez, A., Ortega-Cisneros, S., Pinedo-Díaz, G., García-Contreras, M. S., Alvarado-Castillo, B., Moya-Sánchez, E. U.: Automatic cropping of retinal fundus photographs using convolutional neural networks. *Research in Computing Science*, vol. 149, no. 5, pp. 161–167 (2020)
9. Gonzalez, R. C., Woods, R. E.: *Digital image processing*. Pearson Prentice Hall (2001)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778 (2016) doi: 10.1109/CVPR.2016.90
11. Moya-Sánchez, E. U., Xambó-Descamps, S., Sánchez-Pérez, A., Salazar-Colores, S., Cortés, U.: A trainable monogenic ConvNet layer robust in front of large contrast changes in image classification. *IEEE access*, vol. 9, pp. 163735–163746 (2021) doi: 10.1109/ACCESS.2021.3128552
12. Pinedo-Díaz, G., Ortega-Cisneros, S., Moya-Sánchez, E. U., Rivera, J., Mejía-Alvarez, P., Rodríguez-Navarrete, F. J., Sánchez, A.: Suitability classification of retinal fundus images for diabetic retinopathy using deep learning. *Electronics*, vol. 11, no. 16 (2022) doi: 10.3390/electronics11162564
13. Porwal, P., Pachade, S., Kokare, M., Deshmukh, G., Son, J., Bae, W., Liu, L., Wang, J., Liu, X., Gao, L., Wu, T., Xiao, J., Wang, F., Yin, B., Wang, Y., Danala, G., He, L., Choi, Y. H., Lee, Y. C., Jung, S. H., et al.: IDRiD: Diabetic retinopathy – segmentation and grading challenge. *Medical Image Analysis*, vol. 59 (2020) doi: 10.1016/j.media.2019.101561
14. Rajpurkar, P., Chen, E., Banerjee, O., Topol, E. J.: AI in health and medicine. *Nature medicine*, vol. 28, no. 1, pp. 31–38 (2022) doi: 10.1038/s41591-021-01614-0
15. Ruamviboonsuk, P., Tiwari, R., Sayres, R., Nganthavee, V., Hemarat, K., Kongprayoon, A., Raman, R., Levinstein, B., Liu, Y., Schaekermann, M., Lee, R., Virmani, S., Widner, K., Chambers, J., Hersch, F., Peng, L., Webster, D. R.: Real-time diabetic retinopathy screening by deep learning in a multisite national screening programme: A prospective interventional cohort study. *The Lancet Digital Health*, vol. 4, no. 4, pp. e235–e244 (2022) doi: 10.1016/s2589-7500(22)00017-6
16. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L. C.: MobileNetV2: Inverted residuals and linear bottlenecks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520 (2018) doi: 10.1109/CVPR.2018.00474
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: *3rd International Conference on Learning Representations (ICLR 2015)* (2015) doi: 10.48550/arXiv.1409.1556

18. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826 (2016) doi: 10.1109/CVPR.2016.308
19. Wei-Ting, D. S., Cheung, C., Lim, G., Wei-Tan, G. S., Quang, N. D., Gan, A., Hamzah, H., Garcia-Franco, R., San-Yeo, I. Y., Lee, S. Y., Wong, E., Sabanayagam, C., Baskaran, M., Ibrahim, F., Tan, N. C., Finkelstein, E. A., Lamoureux, E. L., Wong, I. Y., Bressler, N. M., Sivaprasad, S., et al.: Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*, vol. 318, no. 22, pp. 2211–2223 (2017) doi: 10.1001/jama.2017.18152
20. Zago, G. T., Andreão, R. V., Dorizzi, B., Teatini-Salles, E. O.: Retinal image quality assessment using deep learning. *Computers in Biology and Medicine*, vol. 103, pp. 64–70 (2018) doi: 10.1016/j.compbiomed.2018.10.004