Diagnóstico asistido con algoritmos de clasificación: Caso de retinopatía diabética

Dalia Camacho García-Formentí
Instituto Tecnológico Autónomo de México
Mexico City, Mexico
daliaf172@gmail.com

***Abstract*—En este trabajo exploramos cómo se modifica el desempeño de especialistas en oftalmología al realizar diagnósticos de retinopatía diabética, cuando se le proporciona información adicional de un algoritmo de clasificación. Esta información puede ser simplemente la predicción categórica, la predicción númerica, o una capa interpretable a nivel pixel. Los resultados obtenidos muestran una mejora en el diagnóstico una vez que se cuenta con información adicional. Además, se muestra que en algunos casos el algoritmo de clasificación tiene un mejor desempeño que el de los especialistas.**

***Index Terms*—retinopatía diabética, relación humano-computadora, aprendizaje profundo**

I. Introducción

Los algoritmos de predicción y clasificación son capaces de realizar adecuadamente distintas tareas. Incluso en ámbitos como los de clasificación de reseñas en reales y falsas; identificación de *fake news* [1]; o sistemas de recomendación [2], los algoritmos suelen tener un desempeño superior al que tienen los seres humanos . Sin embargo en el dominio de la medicina no es deseable que el diagnóstico final sea realizado por un algoritmo, especialmente si éste no es interpretable. Ya que los algoritmos pueden encontrar patrones en los datos que no sean un verdadero indicador del comportamiento de la enfermedad como ocurrió en el caso de predicción de neumonía descrito por Caruana [3].

Esto no quiere decir que los algoritmos no puedan ser utilizados por los especialistas para llegar a un mejor diagnóstico. Varios autores como [1], [4]–[6] han evaluado cómo se modifica el desempeño de los seres humanos cuando se apoyan de un algoritmo de predicción. Se ha visto que el desempeño depende de la confianza que se le tiene al algoritmo y ésta, a su vez, depende de qué tan interpretable sea el algoritmo [1], [2], [5]. Esto indica la necesidad de poder comunicar el resultado del algoritmo de una forma interpretable que dé confianza a los especialistas y que además pueda ser auditada con ojo crítico para no caer en una confianza ciega como ocurre en [5].

En el presente trabajo nos enfocamos en la tarea de diagnosticar pacientes con retinopatía diabética referible. Los diagnósticos médicos tienen su verdad de base en la clasificación hecha por especialistas, por lo cual se espera que el desempeño inicial de los oftalmólogos sea alto. Lo cual no ocurre en tareas previamente evaluadas [1], [5] donde se sabe que el desempeño inicial del humano es bajo.

La retinopatía diabética es una enfermedad ocasionada por la diabetes y puede generar ceguera si no se trata oportunamente [7]. La clasificación entre referible y no referible está relacionada con la necesidad de tratamiento [8]. Dada la prevalencia de diabetes a nivel mundial (8.5%) [9], se ha vuelto imperativo el desarrollo de algoritmos que permitan un diagnóstico automático y esto ha sido abordado por [10]–[17]. Estos trabajos se han enfocado principalmente en desarrollar arquitecturas que lleven a mejores diagnósticos, pero no se ha comparado su desempeño con el de los especialistas; no se ha evaluado el desempeño de un diagnóstico en conjunto; y tampoco se han generado sistemas de explicación.

En este trabajo hay tres objetivos principales, los cuales son: comparar el desempeño del algoritmo contra el de los especialistas de forma individual; conocer si el desempeño de los especialistas se modifica al contar con información adicional proveniente del algoritmo; y finalmente definir de qué manera resulta mejor presentar la información del algoritmo.

II. Metodología

A. Bases de datos

Se utilizaron dos bases de datos distintas, una para la fase de entrenamiento de la red y la otra para la fase experimental. Para la fase de entrenamiento se utilizó una base de datos de Kaggle [18] que cuenta con $88,702$ imágenes de fondo de ojo provenientes de EyePACS [19]. Para la segunda fase se utilizó una base de datos proporcionada por la Asociación Para Evitar la Ceguera en México (APEC), la cual cuenta con 215 imágenes de fondo de ojo. La distribución de número de imágenes en cada categoría de RD se muestra en la Tabla I

	EyePACS	**APEC**
RD no referible	71,584	112
RD referible	17,154	103
Total	88,702	215

TABLE I
Distribución de imágenes de EyePACS y APEC.

La base de EyePACS contaba con un etiquetado de RD correspondiente a 5 fases en las que puede presentarse la enfermedad, posteriormente fueron condensadas en las categorías de referible y no referible, de acuerdo con [8].

El etiquetado de las imágenes de APEC fue realizado por dos expertas en retina de APEC, aunque el diagnóstico que realizaron fue para cada una de las cinco categorías únicamente corroboramos que coincidieran en si era referible o no referible. Hubo 28 imágenes en las que no se dio este

acuerdo y un tercer experto realizó el diagnóstico para estas imágenes. Además, evaluaron la calidad de las imágenes y definieron que 30 de ellas eran de mala calidad.

B. Arquitectura

En este trabajo se utilizó una arquitectura basada en la red $VGG16$ [20]. Esta arquitectura no realiza explícitamente la localización de lesiones para el diagnóstico de RD referible. Por lo tanto no se necesita el etiquetado previo de las lesiones, el cual debe ser realizdo por especialistas.

La arquitecutra VGG16 es una red convolucional que recibe imágenes de tamaño 224×224 en formato RGB y cuenta con 16 capas con pesos a entrenar. Los filtros utilizados en las capas convolucionales son de tamaño 3×3, ya que éste es el tamaño mínimo requerido para capturar las nociones básicas de dirección tales como arriba, abajo, izquierda, derecha y centro. En las capas convolucionales se utiliza un *stride* de 1 y un *padding* de 1. Además, se tienen 5 capas de agregación *max-pooling*, el cual se hace con ventanas de 2×2 y un *stride* de 2. Finalmente la red VGG16 tiene 3 capas totalmente conexas; las primeras dos con $4,096$ unidades y activación $ReLU$ y la última capa de tamaño $1,000$ con activación *softmax* [20].

En la arquitectura utilizada en el presente trabajo se eliminaron las capas densas y fueron reemplazadas por una capa de *drop out* con probabilidad de 0.5, una capa densa con $1,024$ unidades y activación ReLU, seguida por una capa de *drop out* con probabilidad de 0.5 y una capa densa con una unidad y activación sigmoidal.

La muestra de EyePACS se dividió en conjunto de entrenamiento y validación y conjunto de prueba siguiendo lo realizado por Voets *et al.* en [17], por lo que el conjunto de entrenamiento y validación tienen $57,146$ imágenes las cuales se dividen 80% en entrenamiento y 20% en validación; y el conjuto de prueba tiene $8,706$ imágenes.

La red se entrenó utilizando el optimizador Adam con la librería Keras [21] en Python 3.5 [22]. Los pesos en las capas convolucionales se inicializaron con los de la red preentrenada para imágenes de ImageNet [23] y los de las capas densas se incializaron de forma aleatoria. Se definió un término temprano del entrenamiento si después de 10 épocas la función de pérdida evaluada sobre el conjunto de validación no mejoraba y se guardó el mejor modelo en el conjunto de validación. La red se entrenó con distintos metaparámetros y se seleccionó la combinación con mejor desempeño en el conjunto de prueba de EyePACS. Las tres redes cuya predicción promedio dio los mejores resultados fueron entrenadas con las características de la Tabla II.

Regularización	**Semilla numpy**	**Semilla tensorflow**	**Tasa de Aprendizaje**	**Tamaño de Lote**
Ridge con $\lambda = 0.01$	1	2	$1e^{-5}$	16
Ridge con $\lambda = 0.01$	958	384	$1e^{-5}$	32
Ridge con $\lambda = 0.02$	1	2	$1e^{-5}$	16

TABLE II
METAPARÁMETROS DE LOS 3 MEJORES MODELOS

El área bajo la curva característica de operación (ROC-AUC) de los 3 modelos y del ensamble para el conjunto de prueba de EyePACS se muestra en la tabla III. Las curvas ROC del ensamble de modelos para los datos de EyePACS y APEC se muestran en la Figura 1.

Modelo	**ROC-AUC**
1	0.974
2	0.960
3	0.969
Ensamble	0.977

TABLE III
ÁREA BAJO LA CURVA ROC EN EYEPACS PARA LOS TRES MODELOS UTILIZADOS Y EL ENSAMBLE

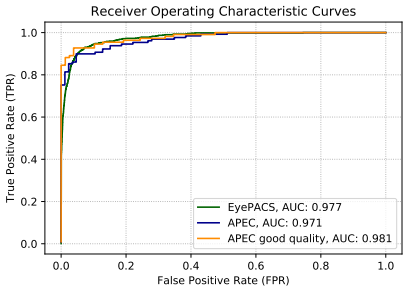

Fig. 1. Curvas ROC para el ensamble de modelos en EyePACS y APEC.

Se puede observar que el desempeño del modelo ensamblado con los datos de APEC es bastante bueno con un AUC de 0.971 si se consideran todas las imágenes y un 0.981 si se consideran únicamente las imágenes de buena calidad. Esto, además de dar mayor confianza de su uso en datos nuevos, indica que la RD se expresa de forma similar en ambas poblaciones y que la clasificación que realizan los oftalmólogos de esta enfermedad es consistente en ambos casos.

C. Selección del modelo explicativo

La selección de un modelo explicativo tuvo como finalidad encontrar una representación de la explicación que pudiera ser interpretable, relevante y visualmente clara. Para esto exploramos el funcionamiento de LIME [24], una explicación *ad-hoc* y el funcionamiento de LRP [25]. Además, consideramos si la representación debía estar basada en la selección de segmentos, utilizando el algoritmo SLIC [26], o en mapas de calor.

Aunque LIME lograba captar las lesiones de forma correcta en la mayoría de los casos con RD referible, las lesiones que encontraba eran dependientes del número de iteraciones. En otras ocasiones seleccionaba partes del fondo como relevantes para un diagnóstico de RD referible. Al analizar el funcionamiento de LIME y la forma en que se expresa la RD vimos que bloquear el 50% de la imagen (que es la proporción de segmentos que selecciona LIME por *default*) no necesariamente se modifica el diagnóstico, ya que las lesiones aparecen en todo el ojo. Y para poder captar de forma exacta la importancia de cada segmento hubiera sido necesario

probar exahustivamente todas las posibles combinaciones de segmentos. Lo cual es computacionalmente poco eficiente.

Por lo cual se propuso un método que también bloquea partes de la imagen, para encontrar la importancia de cada pixel y así generar mapas de calor.

Para obtener los mapas de calor se definió una ventana deslizante de tamaño 64×64 con tamaño de paso de 4 y a la imagen original se le agregó un borde negro de tamaño 63 en cada dirección. En cada iteración los pixeles fuera de la ventana se colorearon de negro y se obtuvo la predicción correspondiente, la importancia que se le atribuyó a cada pixel fue dada por el promedio de las predicciones en las cuales el pixel estuvo dentro de la ventana. Para suavizar las transiciones de color dentro del mapa se realizó una convolución con kernels de tamaño 8×8 cuyo valor en cada entrada fue $\frac{1}{64}$.

Los mapas de calor construidos con este método logran cubrir la mayor parte de las lesiones, pero al igual que LIME se hacen muchas evaluaciones del modelo en distintas imágenes. Esto no sólo es costoso computacionalmente, si no que no es posible saber, qué fue lo que llevo a una u otra predicción. Es por eso que se buscaron métodos que caen en los métodos de representación, según la taxonomía de Gilpin *et al.* [27].

Finalmente se utilizó el método LRP [25], que es un método de explicación por representación del rol de las capas, en este caso sobre los pixeles, los cuales corresponden a la capa de entrada. Inicialmente se probó el método en su versión original, pero esto hizo diverger mucho algunos valores dependiendo del modelo de entrada. Además, éste método considera el impacto de los pesos negativos, los cuales al propagarse hasta la capa de etrada señalan las partes de la imagen que son evidencia en contra de la predicción. En el contexto de RD la evidencia negativa no es relevante, ya que no se considera que existan indicadores de un ojo sano, fuera de la ausencia de lesiones. Por esta razón se optó por el método LRP $\alpha - \beta$, el cual divide los pesos positivos y negativos. Esto evita que los denominadores se anulen o tomen valores muy pequeños haciendo explotar las importancias [28]. Por otro lado, si se selecciona $\alpha = 1$ y $\beta = 0$ únicamente se consideran las unidades que contribuyen de forma positiva al diagnóstico de RD.

D. Diseño de Experimento

La finalidad del experimento fue evaluar y comparar el desempeño de expertos en retina al realizar un diagnóstico de RD en las categorías de RD referible y RD no referible ante distintos niveles de información de un algoritmo de clasificación.

Para lograr esto contamos con la participación de 17 estudiantes de la especialidad en retina en APEC, en su mayoría del tercer año de especialidad. Los participantes realizaron las tareas del experimento de forma voluntaria y dieron su consentimiento al ingresar a la plataforma web donde se montó el experimento, la cual fue la plataforma web Empirica [29] cuyo propósito es poder llevar a cabo y monitorear experimentos en internet.

Cada participante realizó el diagnóstico de RD en tres modalidades. La primera fue hacer el diagnóstico únicamente con la imagen original de fondo de ojo. Otra modalidad fue mostrarle a los participantes la imagen original y la predicción binaria del algoritmo con la leyenda “Referable” o “Non Referable”. La tercera modalidad consistió en mostrarle a los participantes la imagen original, el mapa de calor y la predicción numérica del algoritmo en una escala entre -100 y 100, donde -100 quería decir que la evaluación del algoritmo fue 0, por lo tanto no referible y 100 que el algoritmo dio 1, por lo tanto referible. Los participantes utilizaron esta misma escala para mostrar qué tan seguros estaban de su diagnóstico. Las últimas dos modalidades se presentaron en orden aleatorio, para evitar que el orden de éstas influyera en los resultados.

En cada una de las etapas los participantes evalúan 15 imágenes seleccionadas de forma aleatoria entre 100 imágenes elegidas para el experimento. Estas 100 imágenes fueron elegidas de la base de datos de APEC de la siguiente manera. Se eliminaron las 30 de baja calidad; de las 185 restantes se seleccionó una única imagen por paciente, con lo que quedaron 110 imágenes distintas; finalmente de estas 110 se eligieron de forma aleatoria 50 con RD no referible y 50 con RD referible.

Al final del experimento se les preguntó el año que cursaban en la especialidad, si fue útil tener la respuesta binaria del algoritmo, si fue útil tener la explicación del algoritmo (respuesta numérica y mapa de calor) y si tenían comentarios adicionales.

III. Resultados

El desempeño de los participantes se evaluó en términos de la exactitud, es decir en términos de la proporción de respuestas correctas. En la tabla IV se muestra la exactitud del algoritmo y de los estudiantes en cada fase. También se muestra la mejora porcentual que tienen los médicos en su desmpeño cuando observan la predicción del algoritmo y los mapas de calor. Estos mismos resultados se pueden observar en la Figura 2.

Modalidad	Algoritmo	Médicos	Mejora porcentual
Todas	0.92 (0.90, 0.925)	0.82 (0.80, 0.85)	-
Solitario	0.93 (0.91, 0.96)	0.87 (0.83, 0.91)	-
Respuesta binaria	0.92 (0.89, 0.95)	0.87 (0.83, 0.90)	5.53 (2.91, 8.02)
Respuesta numérica y mapa de calor	0.89 (0.85, 0.92)	0.79 (0.75, 0.83)	7.39 (2.91, 12.46)

TABLE IV
Proporción de respuestas correctas en cada modalidad del experimento

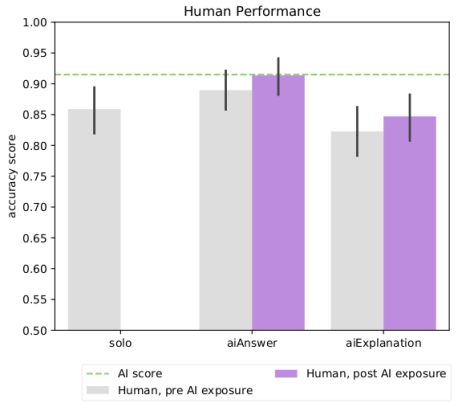

Fig. 2. Desempeño de los médicos en las distintas modalidades del experimento

En general se puede ver que en promedio el algoritmo tuvo un mejor desempeño que los médicos. Sin embargo, si consideramos la fase en solitario como el desempeño de base, no necesariamente es concluyente que éste sea siempre el caso.

En la Figura 3 se muestra el desempeño de los médicos antes de ver la respuesta del algoritmo contra el desempeño después de haber visto la respuesta binaria o el mapa de calor y la respuesta numérica. Las líneas verdes muestran el desempeño del algoritmo.

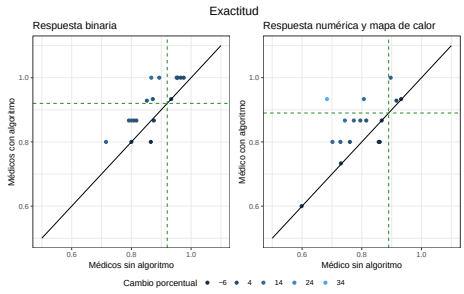

Fig. 3. Caption

En cuanto al desempeño de los médicos, se observa una mejora cuando se les presenta información adicional provista por el algoritmo. En el caso de la respuesta binaria hubo una mejora porcentual de 5.53% respecto al desempeño inicial, mientras que el mostrar el mapa de calor y la respuesta numérica indica una mejora porcentual del 7.39% sobre el desempeño inicial. Esto indica que los médicos confiaron en el algoritmo para cambiar algunas de sus predicciones y que esto tuvo un impacto positivo en su desempeño en la mayoría de los casos. De hecho, únicamente dos médicos empeoraron su desempeño en cada caso. Hubo dos médicos en respuesta binaria y uno en respuesta numérica y mapas de calor que no cambiaron su desempeño. En todos los demás casos hubo una mejora en el desempeño.

IV. Conclusiones

A partir del experimento se pueden hacer dos conclusiones fundamentales. La primera es que el algoritmo que se utilizó produce diagnósticos con una exactitud similar a la de los médicos. Esto quiere decir que ante una posible masificación del diagnóstico el algoritmo es una herramienta confiable.

En segundo lugar el desempeño de los médicos mejora cuando tienen información adicional del algoritmo. Además, al mostrar los mapas de calor éste deja de ser una caja negra. Incluso, los médicos pueden saber qué lesiones está encontrando y cuáles no está tomando en cuenta.

A pesar de que en promedio mejoró más el desempeño con el mapa de calor y la respuesta numérica, no es contundente que sea la mejor forma de compartir la información del algoritmo con los médicos. Probablemente una combinación entre la respuesta binaria y el mapa de calor dé mejores resultados. A pesar de esto contar con algún tipo de información adicional produce mejores resultados.

A partir de estos resultados se puede buscar llevar a la práctica tanto la masificación del diagnóstico como el diagnóstico asistido. No solamente para RD referible y no referible; si no ampliarlo a las 5 categorías, ya que la urgencia del tratamiento depende de la categoría en que se encuentre el paciente en caso de que haya una saturación del sistema de salud. Más aún esto puede extenderse a otras enfermedades. En el caso de enfermedades de la retina, como edema macular, se podrían aprovechar los pesos de las redes que se entrenaron para hacer *transfer learning* si se cuenta con pocos datos.

Agradecimientos

Para poder realizar este trabajo se contó con el apoyo del Dr. Alejandro Noriega Campero, el Dr. Abdullah Almaatouq, Rami Manna y Houssam Kherraz de Massachusetts Institute of Technology; el apoyo de la Dra. Daniela Meizner de la Asociación Para Evitar la Ceguera; de la Asociación Para Evitar la Ceguera y del Consejo Nacional de Ciencia y Tecnología.

References

[1] V. Lai and C. Tan, "On human predictions with explanations and predictions of machine learning models: A case study on deception detection," *CoRR*, vol. abs/1811.07901, 2018. http://arxiv.org/abs/1811.07901.

[2] M. Yeomans, A. Shah, S. Mullainathan, and J. Kleinberg, "Making sense of recommendations," *Journal of Behavioral Decision Making*, pp. 1–12, 2019.

[3] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, (New York, NY, USA), pp. 1721–1730, ACM, 2015. http://doi.acm.org/10.1145/2783258.2788613.

[4] J. M. Logg, J. A. Minson, and D. A. Moore, "Algorithm appreciation: People prefer algorithmic to human judgment," *Organizational Behavior and Human Decision Processes*, vol. 151, no. C, pp. 90–103, 2019.

[5] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Vaughan, and H. M. Wallach, "Manipulating and measuring model interpretability," *CoRR*, vol. abs/1802.07810, 2018. http://arxiv.org/abs/1802.07810.

[6] M. Yin, J. Wortman Vaughan, and H. Wallach, "Understanding the effect of accuracy on trust in machine learning models," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, (New York, NY, USA), pp. 279:1–279:12, ACM, 2019. http://doi.acm.org/10.1145/3290605.3300509.

[7] D. Aliseda and L. Berástegui, "Retinopatía Diabética," *Anales del Sistema Sanitario de Navarra*, vol. 31, pp. 23 – 34, 00 2008. http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1137-66272008000600003&nrm=iso.

[8] M. D. Abràmoff, J. C. Folk, D. P. Han, J. D. Walker, D. F. Williams, S. R. Russell, P. Massin, B. Cochener, P. Gain, L. Tang, M. Lamard, D. C. Moga, G. Quellec, and M. Niemeijer, "Automated Analysis of Retinal Images for Detection of Referable Diabetic RetinopathyAutomated Analysis of Retinal Images," *JAMA Ophthalmology*, vol. 131, pp. 351–357, 03 2013.

[9] H. O. World, "Diabetes," Oct 2018. https://www.who.int/es/news-room/fact-sheets/detail/diabetes.

[10] M. Bhaskaranand, C. Ramachandra, S. Bhat, J. Cuadros, M. G. Nittala, S. Sadda, and K. Solanki, "Automated diabetic retinopathy screening and monitoring using retinal fundus image analysis," *Journal of Diabetes Science and Technology*, vol. 10, no. 2, pp. 254–261, 2016. PMID: 26888972, https://doi.org/10.1177/1932296816628546.

[11] Y. Yang, T. Li, W. Li, H. Wu, W. Fan, and W. Zhang, "Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks," *CoRR*, vol. abs/1705.00771, 2017. http://arxiv.org/abs/1705.00771.

[12] C. Kaya, O. Erkaymaz, O. Ayar, and M. Özer, "Impact of hybrid neural network on the early diagnosis of diabetic retinopathy disease from video-oculography signals," *Chaos, Solitons & Fractals*, vol. 114, pp. 164–174, 2018. http://www.sciencedirect.com/science/article/pii/S0960077918304909.

[13] Z. Wang, Y. Yin, J. Shi, W. Fang, H. Li, and X. Wang, "Zoom-in-net: Deep mining lesions for diabetic retinopathy detection," *CoRR*, vol. abs/1706.04372, 2017. http://arxiv.org/abs/1706.04372.

[14] R. F. Mansour, "Deep-learning-based automatic computer-aided diagnosis system for diabetic retinopathy," *Biomedical Engineering Letters*, vol. 8, pp. 41–57, Feb 2018.

[15] G. Quellec, K. Charrière, Y. Boudi, B. Cochener, and M. Lamard, "Deep image mining for diabetic retinopathy screening," *Medical Image Analysis*, vol. 39, pp. 178 – 193, 2017. http://www.sciencedirect.com/science/article/pii/S136184151730066X.

[16] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," *JAMA*, vol. 316, pp. 2402–2410, 12 2016. https://doi.org/10.1001/jama.2016.17216.

[17] M. Voets, K. Mollersen, and L. A. Bongo, "Replication study: Development and validation of deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *CoRR*, vol. abs/1803.04337, 2018. http://arxiv.org/abs/1803.04337.

[18] Kaggle, "Diabetic retinopathy detection." https://www.kaggle.com/c/diabetic-retinopathy-detection/data.

[19] M. Bhaskaranand, J. Cuadros, E. Ramachandra, S. Bhat, M. G. Nittala, S. R. Sadda, and K. Solanki, "Eyeart + eyepacs: Automated retinal image analysis for diabetic retinopathy screening in a telemedicine system," in *Chen X, Garvin MK, Liu JJ, Trusso E, Xu Y editors. Proceedings of the Ophthalmic Medical Image Analysis Second International Workshop, OMIA 2015, Held in Conjunction with MICCAI 2015*, pp. 105–112, 2015. https://doi.org/10.17077/omia.1033.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014.

[21] F. Chollet *et al.*, "Keras." https://keras.io, 2015.

[22] G. Van Rossum and F. L. Drake Jr, *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009. http://www.image-net.org/papers/imagenet_cvpr09.bib.

[24] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," *CoRR*, vol. abs/1602.04938, 2016. http://arxiv.org/abs/1602.04938.

[25] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," 2017.

[26] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 2274–2282, Nov 2012.

[27] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An approach to evaluating interpretability of machine learning," *CoRR*, vol. abs/1806.00069, 2018. http://arxiv.org/abs/1806.00069.

[28] W. Samek, G. Montavon, A. Binder, S. Lapuschkin, and K.-R. Müller, "Interpreting the Predictions of Complex ML Models by Layer-wise Relevance Propagation," *arXiv e-prints*, p. arXiv:1611.08191, Nov 2016.

[29] A. Almaatouq and N. Paton, "Empirica · easy multiplayer interactive experiments in the browser."