tvst

Special Issue

Automated Quality Assessment and Image Selection of Ultra-Widefield Fluorescein Angiography Images through Deep Learning

Henry H. Li[1,2], Joseph R. Abraham[1], Duriye Damla Sevgi[1], Sunil K. Srivastava[1,3], Jenna M. Hach[1], Jon Whitney[1], Amit Vasanji[4], Jamie L. Reese[1,3], and Justis P. Ehlers[1,3]

[1] The Tony and Leona Campane Center for Excellence in Image-Guided Surgery and Advanced Imaging Research, Cole Eye Institute, Cleveland Clinic, Cleveland, OH, USA
[2] School of Medicine, Case Western Reserve University, Cleveland, OH, USA
[3] Vitreoretinal Service, Cole Eye Institute, Cleveland Clinic, Cleveland, OH, USA
[4] ERT, Cleveland, OH, USA

Correspondence: Justis P. Ehlers, Cole Eye Institute, Cleveland Clinic, 2022 East 105th Street, I Building, Cleveland, OH 44106, USA. e-mail: ehlersj@ccf.org

Received: April 16, 2020
Accepted: July 21, 2020
Published: September 17, 2020

Keywords: fluorescein angiography; retinal vasculature; diabetic retinopathy; retinal blood flow

Citation: Li HH, Abraham JR, Sevgi DD, Srivastava SK, Hach JM, Whitney J, Vasanji A, Reese JL, Ehlers JP. Automated quality assessment and image selection of ultra-widefield fluorescein angiography images through deep learning. Trans Vis Sci Tech. 2020;9(2):52, https://doi.org/10.1167/tvst.9.2.52

Purpose: Numerous angiographic images with high variability in quality are obtained during each ultra-widefield fluorescein angiography (UWFA) acquisition session. This study evaluated the feasibility of an automated system for image quality classification and selection using deep learning.

Methods: The training set was comprised of 3543 UWFA images. Ground-truth image quality was assessed by expert image review and classified into one of four categories (ungradable, poor, good, or best) based on contrast, field of view, media opacity, and obscuration from external features. Two test sets, including randomly selected 392 images separated from the training set and an independent balanced image set composed of 50 ungradable/poor and 50 good/best images, assessed the model performance and bias.

Results: In the randomly selected and balanced test sets, the automated quality assessment system showed overall accuracy of 89.0% and 94.0% for distinguishing between gradable and ungradable images, with sensitivity of 90.5% and 98.6% and specificity of 87.0% and 81.5%, respectively. The receiver operating characteristic curve measuring performance of two-class classification (ungradable and gradable) had an area under the curve of 0.920 in the randomly selected set and 0.980 in the balanced set.

Conclusions: A deep learning classification model demonstrates the feasibility of automatic classification of UWFA image quality. Clinical application of this system might greatly reduce manual image grading workload, allow quality-based image presentation to clinicians, and provide near-instantaneous feedback on image quality during image acquisition for photographers.

Translational Relevance: The UWFA image quality classification tool may significantly reduce manual grading for clinical- and research-related work, providing instantaneous and reliable feedback on image quality.

Introduction

Fluorescein angiography is a critical tool in the diagnosis and management of retinal disease, such as diabetic retinopathy.[1] In recent years, advancements in angiographic imaging technology have enabled ultra-widefield fluorescein angiography (UWFA) capable of capturing 200° fields of view and visualizing up to 3.2 times more retinal area compared with conventional imaging.[2,3] More of the retina can be imaged with a single image, which often translates to greater ease for the patient and photographer.[4] In addition to being more time consuming, conventional angiography also

Table 1. Grading Criteria

	Field of View	Optic Disc/Macula Visualization	Contrast	Macular Centering
Ungradable	<50%	Poor visualization	Poor contrast	Optic disc and macula may be off-centered
Poor	50%+	Moderate blurring	Lower contrast	Optic disc and macula may be off-centered
Good	70%+	Visible with slight blurring	Moderate contrast	Optic disc and macula may be slightly off-centered
Best	90%+	Fully visible without blurring	Great contrast throughout	Optic disc and macula are centered

suffers from variable image quality. According to one study, 31.6% of conventional angiographic images were ungradable, primarily due to media opacities and poor eye dilation.[5]

Although eye dilation is less of a factor in non-mydriatic UWFA systems, UWFA is similarly affected by media opacities (e.g., vitreous debris, hemorrhage, cataract), lid artifacts, optimal eye–camera distance, sufficient dye infusion, injection-to-image time, and centration.[6] Due to the variable quality, typical sample sizes range from 20 to 50 images to ensure that sufficient numbers of good-quality images are obtained. Following acquisition, physicians must manually review this large quantity of images. This time-consuming process can significantly limit work-flow efficiency, particularly in busy retina clinics, and can reduce the time available to review the optimal images. Moreover, if it is discovered that no images of sufficient quality were obtained, it is likely that the patient is no longer at the camera, thus requiring an additional angiography study. Finally, significant human reader time is required for reading centers and clinical trials to identify images of optimal quality for review.

Consideration of image quality is an integral step toward obtaining high-value clinical diagnostics. Previous studies on image quality in other imaging modalities, such as optical coherence tomography angiography (OCTA), have demonstrated significant impacts on measurements made by automated segmentation and analysis software when image quality was reduced.[7] Because UWFA images are often highly complex, interpretation errors can be propagated when a reliable image quality assessment is lacking.

An automated quality assessment system could dramatically improve workflow, enhance physician interpretation efficiency, optimize image acquisition, and enable automated computational image analysis. In the short term, such a tool could provide immediate feedback to photographers during image acquisition to enable repeat imaging. Deep learning convolutional models have significantly enhanced the ability to segment medical images for optical applications, as well as address complex classification tasks.[8,9] These qualities make them an excellent candidate for automating image quality verification and feedback.

Methods

Images and Grading

UFWA images previously obtained during clinical assessment of retinal disease were used to create the datasets utilized for this study. Institutional review board approval was obtained as part of a retrospective UWFA image analysis study. UWFA images were acquired from both the Optos 200Tx and California imaging systems (Optos plc, Dunfermline, Scotland). A total of 3543 images were selected for the training set for this model. The test set was composed of 392 independent images for evaluating model performance and inter-reader agreement. This quality assessment was diagnosis and UWFA-indication agnostic.

Ground-truth image quality for images used in training was assessed by an image analyst review and classified into one of four categories (ungradable, poor, good, or best) based on key factors such as field of view, obscuration from external features, contrast, and centering (Table 1). Best images had at most one minor quality issue, such as the presence of eyelashes with greater than 90% field of view. Good images had at most two minor or one moderate quality issue, such as slight optic disc and macula centration issues, or greater than 70% field of view. Poor images had at most two moderate or one significant quality issues, such as poor optic disc and macula centration, or greater

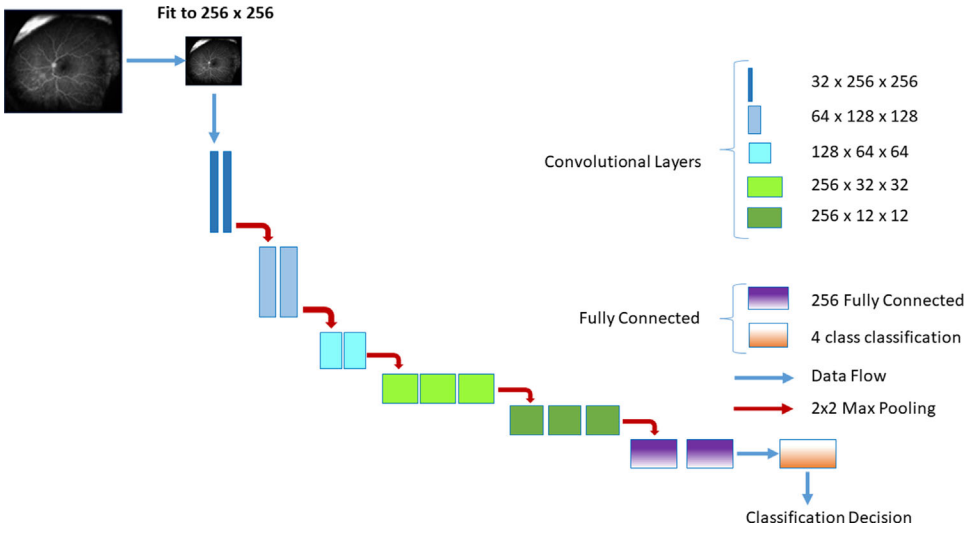

Figure 1. Convolutional model architecture.

than 50% field of view. Specific criteria for ungradable quality included complete obstruction of the optic disc or macula, poor dye visibility, and highly restricted field of view less than 50%.

The testing set was manually graded by two trained image analysts (HL, JRA) with disagreements resolved by a third independent expert reader (DDS). Inter-reader reliability was calculated by taking the number of pairwise comparisons in agreement between the first and second reader over the total number of pairwise comparisons. In addition, a balanced set across quality categories was created using manually graded images of the eyes that were not included in the training set.

Machine Learning Training

The convolutional model is similar to many models used previously, using the encoding layers of a U-net style model, with fully connected layers at the end for classification.[10] This convolutional model was fed into a two-layer fully connected network before final classification. The model architecture is illustrated in Figure 1. This model also uses a 5 × 5 convolutional kernel, which allows for slightly more contextual information to be used, which has been shown to improve performance.[11] The deep learning model was trained on a Quadro M620 graphics card (NVIDIA, Santa Clara, CA), using fivefold cross validation with 80% training, 10% validation, and 10% internal testing separate from both independent training sets. The selected model was the model with the best performance on the internal test set. The training loss function was binary cross entropy, with an Adam optimizer and a loss rate of 1e-4. (Fig. 1)

Model Testing and Performance

Automated classification was completed on a testing set of 392 images to assess model performance. Images were automatically sorted into quality categories of ungradable, poor, good, and best. Machine learning performance was determined by calculating the sensitivity, specificity, and accuracy of the model when evaluated on testing sets. Performance was also assessed via a receiver operating characteristic (ROC) curve. Because ROC curves are restricted to two-classifier problems, performance was reported using the class separation between ungradable and gradable images. A balanced set of 100 images independent from the training set was used to evaluate the potential bias of the unbalanced training data on the machine learning model. This balanced set included 50 ungradable/poor images and 50 good/best images.

Results

Image Classification

A total of 3935 clinical UWFA images were graded by an expert reader to establish ground-truth image quality. For the 3935 images used in this study, 1627 (41.3%) were ungradable, 1156 (29.4%) were poor, 1042 (26.5%) were good, and 110 (2.8%) were best. Manual and automated grading distributions of the 392 testing set images are presented in Table 2, which shows the consistency of grading distributions during automated classification. Inter-reader reliability was calculated to be 84.2%, which

Table 2. Grading Distribution in Testing Set Images

	Testing Set, *n* (%)	
	Manual Classification	Automated Classification
Ungradable	162 (41.3)	152 (38.8)
Poor	115 (29.3)	117 (29.8)
Good	104 (26.5)	105 (26.8)
Best	11 (2.8)	18 (4.6)

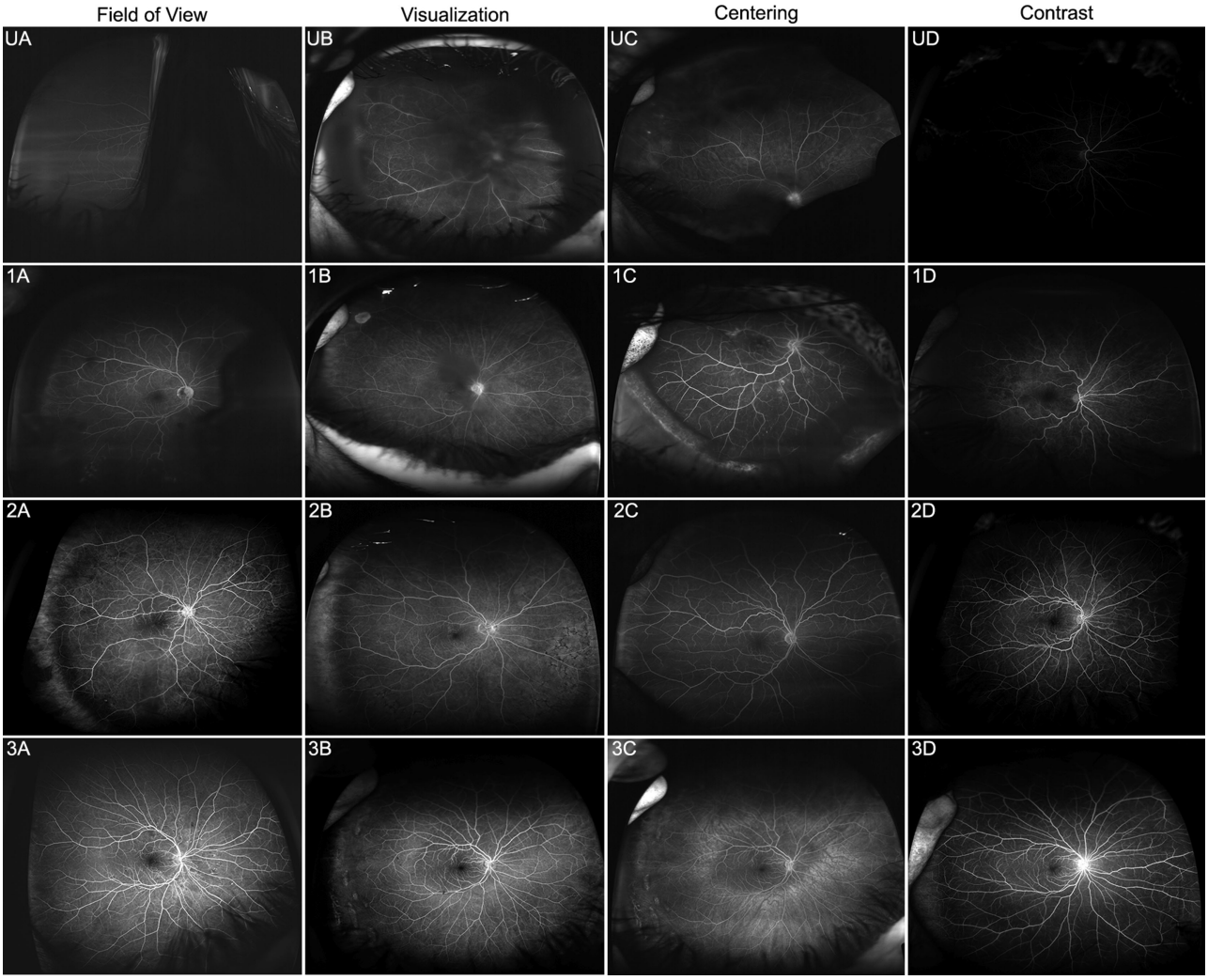

Figure 2. Representative image quality assessment (U, ungradable; 1, poor; 2, good; 3, best). Representative images are selected to demonstrate the quality characteristics of each grade, as determined by the expert image reader. Column A represents images with varying fields of view, column B shows the ranges of visualization of the optic disc/macula region, column C shows the degrees of optic disc centering, and column D shows the various levels of image contrast.

took the pairwise comparisons in agreement between two readers (HL, JRA) over the total number of comparisons. Examples of graded images are provided in Figure 2, showing key factors affecting grade in each category.

Machine Learning Performance

During cross-fold validation, the model reported an average *F*-score performance of 0.74, with a standard deviation of 0.06. The best performing model that

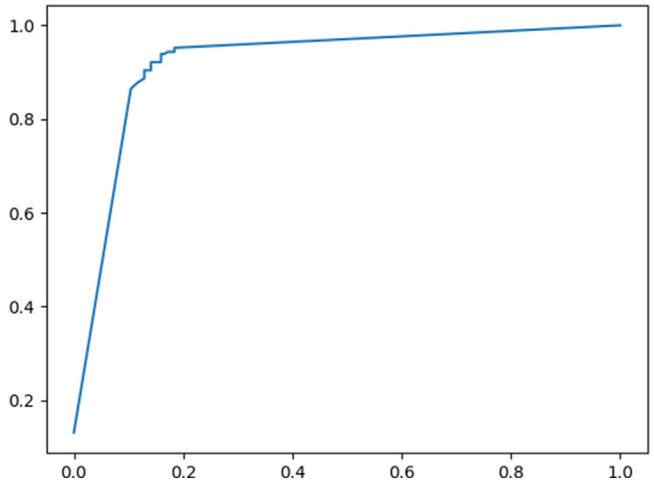

Figure 3. ROC curve for the testing set: ungradable versus gradable images (P, poor; G, good; and B, best).

was used for subsequent analysis had an *F*-score on the internal test set of 0.78. The automated classifier showed a sensitivity of 90.5% and specificity of 87.0% for distinguishing between gradable (poor, good, best) and ungradable images. The automated classifier showed a sensitivity of 78.9% and specificity of 94.1% for distinguishing between optimal quality (good, best) and limited quality (poor, ungradable) images. The overall accuracy of our classifier was calculated to be 89.0% for gradable versus ungradable and 89.3% for recognizing optimal quality versus limited quality. The ROC curve indicates an area under the curve (AUC) of 0.920, measuring performance in a two-class classification between ungradable and gradable images (grades 1–3) (Fig. 3).

When considering eyes that the expert reader graded as ungradable in the test set, 99.4% (161/162) were classified as ungradable or poor by the automatic classifier, demonstrating important consistency at both the high end and low end of the quality spectrum. In the testing set, eight out of 14 best images were classified correctly by our automated tool, whereas six images were classified as good. In addition, 77 out of 99 good images were classified correctly, whereas six were incorrectly classified as best, 15 were classified as poor, and one was classified as ungradable (Table 3). Good-quality images that were classified in lower grades tended to have lower contrast or moderate quality issues in media opacity. The good images incorrectly graded higher generally had two minor issues, such as minor eyelash obscuring the field of view or slightly decreased contrast that prevented expert-based classification as best.

A balanced testing set of 100 images, including 23 ungradable, 27 poor, 31 good, and 19 best images, was sorted using the classifier tool. Assessment of the balanced dataset showed a sensitivity of 98.6%, specificity of 81.5%, and accuracy of 94.0% when differentiating between gradable and ungradable images. For high-quality versus low-quality images, the model showed a sensitivity of 100.0%, specificity of 94.3%, and accuracy of 97.0%. The distribution of grades in automated classification were 27 ungradable, 26 poor, 43 good, and four best images (Table 4). The model is more likely to identify best images as good, which can be attributed to the lower number of best images in the training set (2.8%).

Table 3. Automated Classification Versus Expert Reader

	Expert Reader (*n*)			
Algorithm Assessment	Ungradable	Poor	Good	Best
Ungradable	140	20	1	0
Poor	21	72	15	0
Good	1	25	77	6
Best	0	0	6	8

Table 4. Automated Classification Versus Expert Reader Using a Balanced Testing Set

	Expert Reader (*n*)			
Algorithm Assessment	Ungradable	Poor	Good	Best
Ungradable	22	4	0	1
Poor	1	23	2	0
Good	0	0	27	16
Best	0	0	2	2

Discussion

In this study, a deep learning method was evaluated for the assessment of image quality in UWFA images. The automated quality assessment system achieved a sensitivity of 90.5%, specificity of 87.0%, and overall accuracy of 89.0% in identifying gradable versus ungradable images. In addition, in differentiating between optimal quality (good/best) images versus limited quality (poor/ungradable) images, the automated classifier performed with a sensitivity of 78.9% and specificity of 94.1%. This model can provide rapid image classification-based, clinically relevant image features that can be used to provide near-instantaneous feedback on image quality during acquisition and during the image review process.

Optimization for the deep learning model was centered toward reducing misclassification of gradable

images that were incorrectly classified as ungradable. Although six best images were incorrectly classified as good, the results still demonstrate a low number of false negatives in higher quality images, as none was classified as poor or ungradable. Only 1% (1/99) of the good images were classified as ungradable. The single image classified as ungradable had a high degree of leakage, which may have impacted the classifier tool. Incorrect classifications were heavily biased toward poorer quality images. The complexity and heterogeneity of poor-quality images likely contributed to challenges in accurate automated classification of these images. However, when considering expert-reader-determined ungradable eyes, the automated classifier identified 99.4% (161/162) of these images as either poor or ungradable. This is likely among the most important factors for utilizing a system such as this for automated selection of images for clinician review or quantitative analysis.

Although wide-kernel convolutional models have been used for medical imaging analysis,[12–14] our review of the literature did not reveal that this approach has been applied to UWFA image quality classification. Artifacts such as eyelashes and media opacities, as well as other variations in UWFA image quality, provide a challenge for model adaptation. For this reason, we used a larger training set that was disease agnostic. Previous studies have demonstrated the feasibility of automated image quality classifiers for other modalities, such as fundus photography and OCTA.[15–17]

Interestingly, in en face OCTA, fewer training samples were needed to distinguish superficial vascular structures. The algorithm was trained on 200 OCTA images evaluated by a single image reader and achieved sensitivity, specificity, and accuracy of 90.0% each.[17] Another application of automated image quality classification can be seen with an artificial intelligence fundus image assessment tool recently approved by the Food and Drug Administration. The fundus photograph quality assessment component measures multiple criteria, such as retinal area, focus, and exposures, and then appoints either an adequate or inadequate quality assignment to the image.[15,18] The model was trained on 9963 fundus images and achieved an AUC of 0.978 for predicting gradable retinal fundus images, with sensitivity of 93.9% and specificity of 90.9%.[15] Although there are differences in imaging modality and grading standards when compared with previous work, this study achieved similarly high sensitivity and specificity when identifying gradable versus ungradable images (90.5% and 87.0%, respectively) and when identifying optimal versus limited quality images (78.9% and 94.1%, respectively). Our method achieved an AUC of 0.920 in two-class classification between ungradable and gradable images, demonstrating a large separation between positive and negative classes. Because this algorithm allows the user to adjust operating thresholds, settings can be adjusted to maximize either sensitivity or specificity. This is especially advantageous in clinical settings to minimize false-negative results.

One challenge in developing these systems is the underlying ambiguity in image quality assessment by expert readers, particularly for images that do not neatly fit into a given category. Our UWFA classifier tool outputs four categories of image quality and can also provide binary criteria of gradable and ungradable. Overall, image quality assessment remains a crucial and necessary step to ensure reliable data before identifying disease patterns and characterizing disease progression.[19–21]

There are important limitations to this study that should be acknowledged. The training set contained only 2.8% best images, reflecting the nature of the dataset used and also frequently what is seen clinically in the acquisition of UWFA images. This may have produced bias in the algorithm to sort images to more commonly seen grading categories. Training with a more balanced dataset may decrease bias in future models. Furthermore, adding training data with more pathologic images could help reduce the number of higher pathology images being sorted into lower quality categories. However, the performance of this model on the balanced independent test set was also quite good. The grading system developed for image quality, though detailed and structured, relies on the subjective interpretation of a trained reader which has the potential to introduce important bias. Another potential limitation is that, although explicit features were utilized during the manual selection process (e.g., field of view, contrast) to grade images, the deep learning model may be using other features buried in the images that are unknown to the reader and may ultimately result in unpredictable behaviors in other datasets.[11]

In this report, a platform for quality assessment and image selection of UWFA was developed to optimize clinical imaging management. This tool (1) provides automated image selection for clinical review, (2) can provide rapid real-time feedback to photographers regarding current image quality and enable additional images to be obtained prior to the patient leaving the camera, and (3) may accelerate clinical research by reliably assessing image quality in datasets with numerous images. This is a crucial step that, depending on the image review strategy for the clinician, could take several minutes to achieve but could be reduced to seconds.[7] Further research will include applying this tool to additional datasets, assessing

disease-specific performance, and evaluating phase-specific image selection.

Acknowledgments

Supported by a Research to Prevent Blindness Unrestricted Grant to th Variable definitions of quality standards pathologic makeup both need to be addressed when trying to make objective quality standards.e Cole Eye Institute (RPB1508DM) and by a grant from the National Institutes of Health (NIH K23-EY022947).

Disclosure: **H.H. Li,** None; **J.R. Abraham,** None; **D.D. Sevgi,** None; **S.K. Srivastava,** Regeneron (F), Allergan (F), Gilead (F), Bausch and Lomb (C), Santen (C), Leica (P); **J.M. Hach,** None; **J. Whitney,** None; **A. Vasanji,** ERT (E); **J.L. Reese,** None; **J.P. Ehlers,** Aerpio (C, F), Alcon (C, F), Thrombogenics/Oxurion (C, F), Regeneron (C, F), Genentech (C, F), Novartis (C, F), Allergan (C, F), Roche (C), Leica (C, P), Zeiss (C), Allegro (C), Santen (C)

References

1. Ishibazawa A, Nagaoka T, Takahashi A, et al. Optical coherence tomography angiography in diabetic retinopathy: A prospective pilot study. *Am J Ophthalmol*. 2015;160:35–44.e1.
2. Ehlers JP, Wang K, Vasanji A, Hu M, Srivastava SK. Automated quantitative characterisation of retinal vascular leakage and microaneurysms in ultra-widefield fluorescein angiography. *Br J Ophthalmol*. 2017;101:696–699.
3. Wessel MM, Aaker GD, Parlitsis G, Cho M, D'Amico DJ, Kiss S. Ultra-wide-field angiography improves the detection and classification of diabetic retinopathy. *Retina*. 2012;32:785–791.
4. Kiss S, Berenberg TL. Ultra widefield fundus imaging for diabetic retinopathy. *Curr Diab Rep*. 2014;14:514.
5. Manjunath V, Papastavrou V, Steel D, et al. Wide-field imaging and OCT vs clinical evaluation of patients referred from diabetic retinopathy screening. *Eye (Lond)*. 2015;29:416–423.
6. Mendis KR, Balaratnasingam C, Yu P, et al. Correlation of histologic and clinical images to determine the diagnostic value of fluorescein angiography for studying retinal capillary detail. *Invest Ophthalmol Vis Sci*. 2010;51:5864–5869.
7. Al-Sheikh M, Falavarjani KG, Akil H, Sadda SR. Impact of image quality on OCT angiography based quantitative measurements. *Int J Retinal Vitreous*. 2017;3:13.
8. Fang L, Cunefare D, Wang C, Guymer RH, Li S, Farsiu S. Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search. *Biomed Opt Express*. 2017;8:2732–2744.
9. Loo J, Fang L, Cunefare D, Jaffe GJ, Farsiu S. Deep longitudinal transfer learning-based automatic segmentation of photoreceptor ellipsoid zone defects on optical coherence tomography images of macular telangiectasia type 2. *Biomed Opt Express*. 2018;9:2681–2698.
10. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. *Lect Notes Comput Sci*. 2015;9351:234–241.
11. Peng C, Zhang X, Yu G, Luo G, Sun J. Large kernel matters - improve semantic segmentation by global convolutional network. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*. Piscataway, NJ: Institute of Electrical and Electronics Engineers; 2017:1743–1751.
12. Tennakoon R, Mahapatra D, Roy P, Sedai S, Garnavi R. Image quality classification for DR screening using convolutional neural networks. In: *Proceedings of the Ophthalmic Medical Image Analysis International Workshop*. New York: Springer; 2017:113–120.
13. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In: *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*. 2015 arXiv:1412.7062.
14. Zhang L, Gooya A, Dong B, et al. Automated quality assessment of cardiac MR images using convolutional neural networks. *Lect Notes Comput Sci*. 2016;9968:138–145.
15. Kanagasingam Y, Xiao D, Vignarajan J, Preetham A, Tay-Kearney ML, Mehrotra A. Evaluation of artificial intelligence-based grading of diabetic retinopathy in primary care. *JAMA Netw Open*. 2018;1:e182665.
16. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med*. 2018;1:39.
17. Lauermann JL, Treder M, Alnawaiseh M, Clemens CR, Eter N, Alten F. Automated OCT angiography image quality assessment using a deep

learning algorithm. *Graefes Arch Clin Exp Ophthalmol*. 2019;257:1641–1648.
18. USFDA. FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems [press release]. Available at: https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye. Accessed August 21, 2020.
19. Alsaih K, Lemaitre G, Rastgoo M, Massich J, Sidibé D, Meriaudeau F. Machine learning techniques for diabetic macular edema (DME) classification on SD-OCT images. *BioMed Eng OnLine*. 2017;16:68.
20. Murugeswari S, Sukanesh R. Investigations of severity level measurements for diabetic macular oedema using machine learning algorithms. *Ir J Med Sci*. 2017;186:929–938.
21. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316:2402–2410.