Published in final edited form as:

Ophthalmol Retina. 2019 May ; 3(5): 444–450. doi:10.1016/j.oret.2019.01.015.

Automated Fundus Image Quality Assessment in Retinopathy of Prematurity Using Deep Convolutional Neural Networks

Aaron S. Coyner, BS[1], **Ryan Swan, BS**[1], **J. Peter Campbell, MD**[2], **Susan Ostmo, MS**[2], **James M. Brown, PhD**[3], **Jayashree Kalpathy-Cramer, PhD**[1,3,4], **Sang Jin Kim, MD**[2,5], **Karyn E. Jonas, MD**[6], **R.V. Paul Chan, MD**[6], and **Michael F. Chiang, MD, MA**[1,2] **on behalf of the i-ROP research consortium**

[1]Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, OR

[2]Department of Ophthalmology, Casey Eye Institute, Oregon Health & Science University, Portland, OR

[3]Department of Radiology, Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital/Harvard Medical School, Charlestown, MA

[4]Massachusetts General Hospital & Brigham and Women's Hospital Center for Clinical Data Science, Boston, MA

[5]Department of Ophthalmology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea

[6]Department of Ophthalmology, University of Illinois at Chicago, Chicago, IL

Abstract

Purpose: Accurate image-based ophthalmic diagnosis relies on clarity of fundus images. This has important implications for the quality of ophthalmic diagnoses, and for emerging methods

Address for reprints: Michael F. Chiang, MD, Departments of Ophthalmology & Medical Informatics and Clinical Epidemiology, Casey Eye Institute, Oregon Health & Science University, 3375 SW Terwilliger Boulevard, Portland, OR 97239, Tel: 503-494-7830 | chiangm@ohsu.edu.

Members of the i-ROP research consortium:
Oregon Health & Science University (Portland, OR): Michael F. Chiang, MD, Susan Ostmo, MS, Sang Jin Kim, MD, PhD, Kemal Sonmez, PhD, J. Peter Campbell, MD, MPH. University of Illinois at Chicago (Chicago, IL): RV Paul Chan, MD, Karyn Jonas, RN. Columbia University (New York, NY): Jason Horowitz, MD, Osode Coki, RN, Cheryl-Ann Eccles, RN, Leora Sarna, RN. Weill Cornell Medical College (New York, NY): Anton Orlin, MD. Bascom Palmer Eye Institute (Miami, FL): Audina Berrocal, MD, Catherin Negron, BA. William Beaumont Hospital (Royal Oak, MI): Kimberly Denser, MD, Kristi Cumming, RN, Tammy Osentoski, RN, Tammy Check, RN, Mary Zajechowski, RN. Children's Hospital Los Angeles (Los Angeles, CA): Thomas Lee, MD, Evan Kruger, BA, Kathryn McGovern, MPH. Cedars Sinai Hospital (Los Angeles, CA): Charles Simmons, MD, Raghu Murthy, MD, Sharon Galvis, NNP. LA Biomedical Research Institute (Los Angeles, CA): Jerome Rotter, MD, Ida Chen, PhD, Xiaohui Li, MD, Kent Taylor, PhD, Kaye Roll, RN. Massachusetts General Hospital (Boston, MA): Jayashree Kalpathy-Cramer, PhD; Ken Chang, BS; Andrew Beers, BS. Northeastern University (Boston, MA): Deniz Erdogmus, PhD, Stratis Ioannidis, PhD. Asociacion para Evitar la Ceguera en Mexico (APEC) (Mexico City): Maria Ana Martinez-Castellanos, MD, Samantha Salinas-Longoria, MD, Rafael Romero, MD, Andrea Arriola, MD, Francisco Olguin-Manriquez, MD, Miroslava Meraz-Gutierrez, MD, Carlos M. Dulanto-Reinoso, MD, Cristina Montero-Mendoza, MD.

Portions of this manuscript were presented at the 2018 ARVO annual meeting (Honolulu, HI), and will be presented as a paper presentation at the 2018 American Medical Informatics Association Annual Symposium (San Francisco, CA).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

such as telemedicine and computer-based image analysis. The purpose of this study was to implement a deep convolutional neural network (CNN) for automated assessment of fundus image quality in retinopathy of prematurity (ROP).

Design: Experimental study.

Subjects: Retinal fundus images were collected from preterm infants during routine ROP screenings.

Methods: 6,139 retinal fundus images were collected from 9 academic institutions. Each image was graded for quality (Acceptable Quality [AQ], Possibly Acceptable Quality [PAQ], or Not Acceptable Quality [NAQ]) by three independent experts. Quality was defined as the ability to confidently assess an image for the presence of ROP. Of the 6,139 images, NAQ, PAQ, and AQ images represented 5.6%, 43.6%, and 50.8% of the image set, respectively. Due to low representation of NAQ images in the data set, images labeled NAQ were grouped into the PAQ category, and a binary CNN classifier was trained using 5-fold cross-validation on 4,000 images. A test set of 2,109 images was held out for final model evaluation. Additionally, 30 images were ranked from worst to best quality by six experts via pairwise comparisons and the CNN's ability to rank quality, regardless of quality classification, was assessed.

Main Outcome Measures: CNN performance was evaluated using area under the receiver operating characteristic curve (AUC). A Spearman's rank correlation was calculated to evaluate the overall ability of the CNN to rank images from worst to best quality as compared to experts.

Results: The mean (SD) AUC for 5-fold cross-validation was 0.958 (0.005) for the diagnosis of AQ versus PAQ images. The AUC was 0.965 for the test set. The Spearman's rank correlation coefficient on the set of 30 images was 0.90 as compared to the overall expert consensus ranking.

Conclusions: This model accurately assessed retinal fundus image quality in a comparable manner to that of experts. This fully-automated model has potential for application in clinical settings, telemedicine, and computer-based image analysis in ROP, and for generalizability to other ophthalmic diseases.

Technologies such as digital imaging, telemedicine, and artificial intelligence for image analysis are beginning to revolutionize the practice of ophthalmology.[1-11] A critical issue that plagues nearly all medical imaging applications is poor image quality.[12-28] In the best case scenario, poor image quality renders images useless for diagnosis and wastes time and resources due to required follow-up imaging sessions. In the worst case, it leads to incorrect diagnoses, resulting in either over- or under-treatment and the potential for life-altering consequences. To address this issue, we have focused on retinopathy of prematurity (ROP), a potentially-blinding childhood disease.

Advances in medical technology have also been witnessed in the neonatal intensive care unit (NICU).[29] The survival rate of premature infants has dramatically increased over the last few decades.[29] Unfortunately, this has not come without consequences. ROP, a vasoproliferative retinal disease, affects approximately two-thirds of premature infants weighing <1251 grams at birth.[30-33] While ROP has the potential to cause permanent blindness, it is treatable via laser photocoagulation or intravitreal injections of anti-vascular endothelial growth factor (anti-VEGF), if diagnosed promptly.[30] Treatment is initiated for the following retinal findings: Zone I ROP, stages 1, 2 or 3, plus disease present; Zone I ROP, stage 3, plus disease not present, and Zone II ROP, stages 2 or 3, plus disease present.[30] It is obvious that plus disease, defined as “abnormal dilation and tortuosity of the posterior retinal blood vessels in two or more quadrants of the retina,” is a significant indicator of the need for treatment. It is therefore absolutely necessary to diagnose plus disease in an accurate and timely manner. The presence of plus disease in at least two quadrants of the retina is easier to diagnose when image quality is high (Figure 1A). However, as image quality begins to deteriorate, visualization of the retina becomes difficult, if not impossible (Figure 1B,C).

A major barrier to timely ROP treatment is a lack of access to ROP experts in both developed and developing countries.[8,15,31-33] Therefore, the implementation of telemedicine and other computer-based image analysis applications that make use of high-quality fundus images is crucial.[11] Recently, we have developed DeepROP, a deep learning model for automated assessment of plus disease in ROP patients.[10] When this model is provided high-quality images, it provides highly accurate diagnoses. However, it is reasonable to assume that images of lower quality will tend to be misclassified more often than images of higher quality. Herein we describe an extension of preliminary work that attempts to address this pitfall – a deep convolutional neural network (CNN) to automatically assess the quality of retinal fundus images.[34] A CNN is an artificial neural network trained to extract features from images. A deep CNN is an extension of this model, which creates new images using the extracted features. Essentially, a deep CNN extracts features from features from features and so on. In the early layers of the network, the extracted features are typically straight lines of various rotations. In the deeper layers of a CNN, features become more abstract. Because there are typically tens of millions of parameters to train (e.g. weights of the edges connecting nodes), we take advantage of a method known as transfer learning. Here, we implement a pretrained CNN architecture, specifically Inception-V3, which has been trained to identify everyday objects, such as cats, cars, trees, dishwashers, etc., and we fine-tune its learned filters for this specific use case.[35,36] This has numerous potential applications, such as a pre-screening method for our ROP diagnostic tool, a quality metric for imaging technicians, or a workflow component for telemedicine-based applications.

METHODS

All data for this study were obtained through the multi-center, NIH-funded, Imaging and Informatics in ROP (i-ROP) study centered at Oregon Health & Science University (OHSU). This study was approved by the institutional review board at the coordinating center (OHSU, Portland, Oregon) and at each of 8 study centers (Columbia University, University of Illinois at Chicago, William Beaumont Hospital, Children’s Hospital Los Angeles, Cedars-Sinai Medical Center, University of Miami, Weill Cornell Medical Center, Asociacion para Evitar la Ceguera en Mexico) and was conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from parents of all infants enrolled.

Retinal Fundus Image Data Sets

Using a RetCam (Natus; Pleasanton, CA), 6,139 wide-angle fundus images were collected from preterm infants during routine ROP screening examinations. Three masked graders evaluated images for ROP stage, zone, plus disease, and image quality (based upon acceptability for diagnosis of ROP). Labels for image quality were: Acceptable Quality (AQ), Possibly Acceptable Quality (PAQ), or Not Acceptable Quality (NAQ). Graders were not told what defined AQ versus PAQ versus NAQ images. The final classification represents a majority vote of the 3 independent assessments of the suitability of an image for the task of ROP classification (zone, stage, and plus). AQ, PAQ, and NAQ images represented 50.8%, 43.6%, and 5.6% of the final data set, respectively (Figure 1). Due to low representation of NAQ images in this data set, NAQ images were combined with images from the PAQ category. The final distribution of the data set was 50.8% AQ images and 49.2% PAQ images. It should be noted, however, that the PAQ label does not necessarily imply that an image is useless for diagnosis, but that a higher-quality image would increase the confidence of the diagnosis being made. For example, it is possible that a diagnosis could be made from an image with half of the retinal image occluded, but that an image grader's confidence might be higher if the entire image were easily visualized.

To assemble the training set, 2,000 AQ images and 2,000 PAQ images were selected at random. These 4,000 images were randomly decomposed into five separate, equally-stratified sets to be used for 5-fold cross-validation. An independent test set was formed using 2,109 randomly selected images that represented the true underlying distribution of AQ to PAQ images. The remaining 30 images were used to create a ranked data set. Briefly, the six experts ranked the smaller set of 30 images from worst quality to best quality. A web-based interface was implemented, which presented each expert with two images and the prompt "Select the higher quality image for the diagnosis of plus disease." After multiple pairwise comparisons, individual expert rankings of worst to best quality images were developed. Using an Elo rating system, all expert rankings were aggregated to form an overall expert consensus ranking of the images.

Model Architecture

This model was built and trained using Keras, a deep learning library for the programming language Python, with the TensorFlow backend (an open source software library for numerical computation using data flow graphs). The convolutional portion of the model made use of a pretrained CNN, specifically Inception-V3.[35] The weights of the CNN were initialized using the values obtained after training the CNN on the ImageNet database, a collection of over 14 million hand-annotated images containing more than 20,000 classes.[36] This reduced training time, as it allowed the CNN to learn basic features of everyday objects by developing filters to extract specific shapes and textures prior to fine-tuning on medical images. Two fully-connected layers were built on top of the convolutional layers. The first layer consisted of 4,096 nodes using a rectified linear unit (ReLU) activation function. Because we sought to discriminate between AQ and PAQ images, the second layer consisted only of a single binary output node. This final layer made use of the sigmoid activation function; images were not only assigned a classification of AQ or PAQ, but the associated probability of belonging to said class was reported. To prevent overfitting, a dropout function with a probability of 0.5 was inserted between the two layers. Inputs to the model were RetCam images of size 640 × 480 × 3 or 1024 × 768 × 3 scaled down to 150 × 150 × 3. All training and test set image pixel values were rescaled into the [0, 1] range. Training set images also had random zoom (± 20%), horizontal flips, and vertical flips applied to them to synthetically increase the size of the training data set and reduce the chance of overfitting.

Model Training and Evaluation

The five subsets of the training set were used to perform 5-fold cross-validation. Briefly, five versions of the CNN were trained and evaluated using unique validation sets and slightly different training sets. Each CNN was evaluated on subset 1, 2, 3, 4, or 5, and trained on the remaining four subsets. This method allows for close approximation of the test error and reduces the probability of overfitting. Training occurred for 100 epochs (iterations). However, the epoch with the lowest validation set error was selected for each of the five CNNs. Training was executed using the following hyperparameters: optimizer: mini-batch gradient descent, batch size: 20, learning rate: 0.001, momentum: 0.9, loss: binary cross-entropy, and validation metric: accuracy. All layers of the model were adjustable (i.e. the convolutional layers were not frozen).

Data Analysis

Following 5-fold cross-validation, the area under the receiver operating characteristics curve (AUC) was computed for each model. The CNN with the highest AUC was selected as the final model, on which test and ranked set predictions were made. Images were input to the CNN, which calculated the probability of an image belonging to the AQ category using the softmax function of the final layer. A score less than 0.5 placed the image into the PAQ category, and a score greater than or equal to 0.5 placed the image into the AQ category. The AUC of the model was evaluated.

As mentioned above, the output of the CNN for any given image was a probability from 0 to 1. These values were used to rank the set of 30 images from worst to best quality for diagnosis of ROP. The Spearman's rank correlation test was used to assess the similarity between the CNN and the consensus ranking of the images by the six experts, as well as the correlation between individual experts.

RESULTS

Classification Performance

The AUCs resulting from 5-fold cross-validation ranged from 0.953 to 0.965, with a mean (SD) of 0.958 (0.005) (Figure 2A). Model 1 was selected as the final model. On the test set, the AUC was 0.965 (Figure 2B), in line with the estimated test set AUC predicted by 5-fold cross-validation (Figure 2A), and the sensitivity and specificity were 93.9% and 83.6%, respectively. Depending upon the application for which the model was implemented, the classification cutoff probability could be increased or decreased to favor sensitivity or specificity (i.e. to avoid false negatives or avoid false positives).

Ranked Set Performance

Figure 3 describes the Spearman's rank correlation coefficients for each individual expert grader's rank, the consensus rank, and the CNN rank. The Spearman's rank correlation test coefficients between experts ranged from 0.89 to 0.97, suggesting a very high correlation of agreement on relative image quality. Unsurprisingly, all experts were highly correlated to the consensus rank (0.94 - 0.97). The correlations between the CNN and individual experts ranged from 0.86 to 0.93, and the correlation between the CNN and the consensus ranking was 0.90, suggesting that the CNN not only has high inter-group discrimination, but high intra-group discrimination. In essence, given two images from the same class, the CNN can recognize which of the two images is of higher quality despite originating from the same lass. This suggests that the model has not only learned the difference between an AQ image or a PAQ image, but that it has learned what features make any retinal fundus image of higher quality than another.

DISCUSSION

We developed a model for the automated assessment of retinal fundus images in retinopathy of prematurity using a deep convolutional neural network. There are two key findings in this study: (1) with a high degree of confidence, the model can distinguish between images of acceptable quality and images of low or questionable quality, and (2) the model ranks image quality similarly to ROP experts, regardless of image quality classification, suggesting that the threshold at which images are classified as AQ or PAQ could be adjusted based upon the model's application.

The use of 5-fold cross-validation allowed us to train multiple models using all available training data while limiting the risk of overfitting. This finding is illustrated in Figure 2A, which shows that all models perform similarly to one another. The aim of cross-validation is to estimate test set performance. The mean (SD) of the five models was 0.958 (0.005). We used the best performing model to assess the independent test set (Figure 2B). The AUC was 0.965, similar to the mean (SD) predicted by 5-fold cross-validation. Taken together, we believe that this model has not overfit the data and that it is highly generalizable to RetCam-acquired ROP images.

An interesting result presented during model assessment on the ranked image data set. When training the CNNs, we cast our problem as a classification task. That is, we only cared to classify images as AQ or PAQ, and were never concerned about the intra-class ordering of images. However, to ensure applicability in use cases where the threshold at which AQ versus PAQ images may be different, it is important for the algorithm to be able rank image quality from worst to best, regardless to which quality class our experts believe an image belongs. In essence, we were testing the ability of the CNN to perform regression, even though it was only trained for classification. On a smaller data set of 30 images, six experts ranked images from worst to best quality for the diagnosis of plus disease via an exhaustive pairwise comparisons process. This provided us with the individual rankings for the image set for each expert, which we were able to combine into an expert consensus ranking. All experts were highly correlated with one another (0.89-0.97) and, unsurprisingly, with the consensus ranking (0.94-0.97; Figure 3). Rather than have the CNN output class labels for each of the 30 images, we collected the probabilities of each image belonging to the AQ class and ordered them from smallest to largest, thereby establishing the CNN's ranking of the 30 images. The CNN was highly correlated to each individual expert (0.86-0.93) and to the consensus ranking (0.90; Figure 3). These results show that our model has a striking ability to rank images, further suggesting that the threshold at which our model classifies images as AQ or PAQ could be adjusted depending upon application.

Overall, these findings demonstrate the robustness of our model: it correctly identifies what our experts consider to be acceptable quality images vs. low and questionable quality images. This study also demonstrates that the threshold at which the model classifies images could be adjusted for other experts or applications. For example, in a telemedicine application where physicians manually review images, the model would likely remain unchanged since it was trained using the opinions of ROP experts. However, implementation as a prescreening method for a computer-based image analysis tool, such as DeepROP, may warrant some modifications.[10] It is possible that a computer-based image analysis tool could still provide a reliable ROP diagnosis on a subset of PAQ images. Therefore, the threshold at which images were binned into the AQ versus PAQ category could be lowered until all images placed into the PAQ category could not be assessed via the computer-based method. [1,2,4-6,8,10]

While we are confident in the model we have trained, there are some limitations. First, only RetCam images were used for training and testing. We did not evaluate model performance on images from other cameras. Differences in field-of-view and lighting could potentially affect the reliability of the model. Recently, ophthalmic lenses for smartphones have been created.[37,38] An interesting area of potential research involves training our model to accurately assess the quality of images acquired from these devices, thereby greatly enhancing the reliability of telemedicine applications. Second, the model was trained using images acquired from premature infants during routine ROP screenings. It is unclear whether this model can accurately classify images acquired from adults or older children with other ocular conditions, and further training of this model with images from those demographics would be beneficial. Third, the model was trained and validated on posterior pole images. In practice, nasal, temporal, superior, and inferior images may be used in addition to posterior pole images for diagnosis of ROP.[30] Further training of this model will include images from various regions of the retina to increase reliability and applicability in true clinical applications. The final limitation of this model is the lack of ability to distinguish a retinal fundus image from images of other items (e.g. non-ophthlamic images). This model was trained as a retinal fundus image quality classifier, not as a general image quality classifier. One could argue that users of this model will only be acquiring and assessing retinal fundus images. But to ensure conformity, a future direction of this work could involve training a CNN to classify images as retinal fundus images or not prior to images being assessed for quality.

We are not the first group to produce a retinal image quality classifier; however, many other classifiers have severe limitations. To the best of our knowledge, Saha et al. have produced the only other retinal image classifier that takes advantage of a CNN.[39] They used AlexNet, an award-winning but older CNN, for assessing the quality of diabetic retinopathy images. Their model performed with an accuracy of 100% on a data set of 3,572 images.[39] However, their image set only included images on which all graders agreed upon the quality of the images (i.e. images without complete agreement were excluded from the test set) which could leave the data with a very bimodal distribution. Furthermore, their data set was severely imbalanced, as only 143 of the 3,572 images were of unacceptable quality.[39] In theory, a naive model (one that only predicts AQ for every image) would be correct 96% of the time. Consequently, it is possible that their CNN would not generalize well in practice. Other groups have implemented linear algorithms for image quality assessment of retinal fundus photos, which have performed well, but all training and test data sets were small in comparison to the data set we used to train, validate, and test our CNN.[13,17,19,27,28] We believe that, because our CNN was rigorously trained on 4,000 images using cross-validation and tested on two separate test sets consisting of 2,109 images and 30 ranked images, it will generalize better and be more robust in practice.

In this study, we implemented a convolutional neural network for the assessment of retinal fundus image quality in retinopathy of prematurity. We have shown that a convolutional neural network is sufficient for providing a high degree of discrimination between acceptable quality and possibly acceptable quality images, and can rank a set of retinal fundus images from worst to best quality. Potential applications of this algorithm range from inclusion in computer-based image analysis pipelines to implementation in fundus cameras, where imaging technicians could be alerted as to whether their captured images were of acceptable quality for diagnosis of disease. More broadly, it should be noted that this methodology is not limited to retinopathy of prematurity or retinal fundus imaging, as it has potential application in different ocular diseases or for different imaging modalities altogether.

Acknowledgments

Financial Support and Disclosures:

Supported by grants T15LM007088, R01EY19474, P30EY010572, K12EY027720, and T32EY23211 from the National Institutes of Health (Bethesda, MD); by unrestricted departmental funding and a Career Development Award (JPC) from Research to Prevent Blindness (New York, NY) ; and by the iNsight Foundation. No funding organizations had any role in the design or conduct of this research. MFC is an unpaid member of the scientific advisory board for Clarity Medical Systems (Pleasanton, CA), a consultant for Novartis (Basel, Switzerland), and an initial member of Inteleretina, LLC (Honolulu, HI). RVPC is a consultant for Visunex Medical Systems (Fremont, CA), Allergan (Irvine, CA), Alcon (Fort Worth, TX), and Bausch and Lomb (St. Louis, MO). JPC and MFC receive research support from Genentech.

REFERENCES

1. Ataer-Cansizoglu E, Bolon-Canedo V, Campbell JP, et al. Computer-Based Image Analysis for Plus Disease Diagnosis in Retinopathy of Prematurity: Performance of the "i-ROP" System and Image Features Associated With Expert Diagnosis. Transl Vis Sci Technol. 2015;4(6):5.
2. Campbell JP, Ataer-Cansizoglu E, Bolon-Canedo V, et al. Expert Diagnosis of Plus Disease in Retinopathy of Prematurity From Computer-Based Image Analysis. JAMA Ophthalmol. 2016;134(6):651–657. [PubMed: 27077667]
3. Castellanos FX, Giedd JN, Marsh WL, et al. Quantitative brain magnetic resonance imaging in attention-deficit hyperactivity disorder. Arch Gen Psychiatry. 1996;53(7):607–616. [PubMed: 8660127]
4. Chiang MF. Image analysis for retinopathy of prematurity: where are we headed? J AAPOS. 2012;16(5):411–412. [PubMed: 23084374]
5. Chiang MF, Gelman R, Martinez-Perez ME, et al. Image analysis for retinopathy of prematurity diagnosis. J AAPOS. 2009;13(5):438–445. [PubMed: 19840720]

6. Chiang MF, Starren J, Du YE, et al. Remote image based retinopathy of prematurity diagnosis: a receiver operating characteristic analysis of accuracy. Br J Ophthalmol. 2006;90(10):1292–1296. [PubMed: 16613919]

7. Lundberg T, Westman G, Hellstrom S, Sandstrom H. Digital imaging and telemedicine as a tool for studying inflammatory conditions in the middle ear--evaluation of image quality and agreement between examiners. Int J Pediatr Otorhinolaryngol. 2008;72(1):73–79. [PubMed: 17983668]

8. Richter GM, Williams SL, Starren J, Flynn JT, Chiang MF. Telemedicine for retinopathy of prematurity diagnosis: evaluation and challenges. Surv Ophthalmol. 2009;54(6):671–685. [PubMed: 19665742]

9. Smith RA, Saslow D, Sawyer KA, et al. American Cancer Society guidelines for breast cancer screening: update 2003. CA Cancer J Clin. 2003;53(3):141–169. [PubMed: 12809408]

10. Brown JM, Campbell JP, Beers A, et al. Automated Diagnosis of Plus Disease in Retinopathy of Prematurity Using Deep Convolutional Neural Networks. JAMA Ophthalmol. 2018;136(7):803–810. [PubMed: 29801159]

11. Quinn GE, Ying GS, Daniel E, et al. Validity of a telemedicine system for the evaluation of acute-phase retinopathy of prematurity. JAMA Ophthalmol. 2014;132(10):1178–1184. [PubMed: 24970095]

12. Bartlett E, DeLorenzo C, Parsey R, Huang C. Noise contamination from PET blood sampling pump: Effects on structural MRI image quality in simultaneous PET/MR studies. Med Phys. 2018;45(2):678–686. [PubMed: 29210075]

13. Bartling H, Wanger P, Martin L. Automated quality evaluation of digital fundus photographs. Acta Ophthalmol. 2009;87(6):643–647. [PubMed: 19719806]

14. Briggs R, Bailey JE, Eddy C, Sun I. A methodologic issue for ophthalmic telemedicine: image quality and its effect on diagnostic accuracy and confidence. J Am Optom Assoc. 1998;69(9):601–605. [PubMed: 9785735]

15. Chiang MF, Wang L, Busuioc M, et al. Telemedical retinopathy of prematurity diagnosis: accuracy, reliability, and image quality. Arch Ophthalmol. 2007;125(11):1531–1538. [PubMed: 17998515]

16. Dietrich TJ, Ulbrich EJ, Zanetti M, Fucentese SF, Pfirrmann CW. PROPELLER technique to improve image quality of MRI of the shoulder. AJR Am J Roentgenol. 2011;197(6):W1093–1100. [PubMed: 22109324]

17. Gajendra Jung Katuwal JK, Ramchandran Rajeev, Sisson Christye, and Rao Navalgund. Automatic Fundus Image Field Detection And Quality Assessment. IEEE Xplore. 2013.

18. Jiang Y, Huo D, Wilson DL. Methods for quantitative image quality evaluation of MRI parallel reconstructions: detection and perceptual difference model. Magn Reson Imaging. 2007;25(5): 712–721. [PubMed: 17540283]

19. Li H, Hu W, Xu ZN. Automatic no-reference image quality assessment. Springerplus. 2016;5(1): 1097. [PubMed: 27468398]

20. Maberley D, Morris A, Hay D, Chang A, Hall L, Mandava N. A comparison of digital retinal image quality among photographers with different levels of training using a non-mydriatic fundus camera. Ophthalmic Epidemiol. 2004;11(3):191–197. [PubMed: 15370551]

21. Niemeijer M, Abramoff MD, van Ginneken B. Image structure clustering for image quality verification of color retina images in diabetic retinopathy screening. Med Image Anal. 2006;10(6): 888–898. [PubMed: 17138215]

22. Patel T, Peppard H, Williams MB. Effects on image quality of a 2D antiscatter grid in x-ray digital breast tomosynthesis: Initial experience using the dual modality (x-ray and molecular) breast tomosynthesis scanner. Med Phys. 2016;43(4):1720. [PubMed: 27036570]

23. Smet MH, Breysem L, Mussen E, Bosmans H, Marshall NW, Cockmartin L. Visual grading analysis of digital neonatal chest phantom X-ray images: Impact of detector type, dose and image processing on image quality. Eur Radiol. 2018.

24. Strauss RW, Krieglstein TR, Priglinger SG, et al. Image quality characteristics of a novel colour scanning digital ophthalmoscope (SDO) compared with fundus photography. Ophthalmic Physiol Opt. 2007;27(6):611–618. [PubMed: 17956367]

25. Takeda H, Minato K, Takahasi T. High quality image oriented telemedicine with multimedia technology. Int J Med Inform. 1999;55(1):23–31. [PubMed: 10471238]

26. Teich S, Al-Rawi W, Heima M, et al. Image quality evaluation of eight complementary metal-oxide semiconductor intraoral digital X-ray sensors. Int Dent J. 2016;66(5):264–271. [PubMed: 27103603]

27. Veiga D, Pereira C, Ferreira M, Goncalves L, Monteiro J. Quality evaluation of digital fundus images through combined measures. J Med Imaging (Bellingham). 2014;1(1):014001. [PubMed: 26158021]

28. Wang S, Jin K, Lu H, Cheng C, Ye J, Qian D. Human Visual System-Based Fundus Image Quality Assessment of Portable Fundus Camera Photographs. IEEE Trans Med Imaging. 2016;35(4): 1046–1055. [PubMed: 26672033]

29. March of Dimes, PMNCH, Save the Children, WHO. Born Too Soon: The Global Action Report on Preterm Birth. 2012.

30. American Academy of Pediatrics, American Academy of Ophthalmology, American Association for Pediatric Ophthalmology and Strabismus. Screening examination of premature infants for retinopathy of prematurity. Pediatrics. 2013;131(1):189–95. [PubMed: 23277315]

31. Blencowe H, Lawn JE, Vazquez T, Fielder A, Gilbert C. Preterm-associated visual impairment and estimates of retinopathy of prematurity at regional and global levels for 2010. Pediatr Res. 2013;74 Suppl 1:35–49. [PubMed: 24366462]

32. Quinn GE. Retinopathy of prematurity blindness worldwide: phenotypes in the third epidemic. Eye Brain. 2016;8:31–36. [PubMed: 28539799]

33. National Eye Institute. Retinopathy of Prematurity. 2018https://nei.nih.gov/health/rop. Accessed August 01, 2018.

34. Coyner AS, Swan R, Brown JM, et al. Deep Learning for Image Quality Assessment of Fundus Images in Retinopathy of Prematurity. AMIA Annual Symposium Proceedings. In press.

35. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. ArXiv e-prints. 2015 https://ui.adsabs.harvard.edu/#abs/2015arXiv151200567S Accessed August 01, 2018.

36. Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge. ArXiv e-prints. 2014https://ui.adsabs.harvard.edu/#abs/2014arXiv1409.0575R. Accessed August 01, 2018.

37. Giardini ME, Livingstone IA, Jordan S, et al. A smartphone based ophthalmoscope. Conf Proc IEEE Eng Med Biol Soc 2014;2014:2177–2180.

38. Wu AR, Fouzdar-Jain S, Suh DW. Comparison Study of Funduscopic Examination Using a Smartphone-Based Digital Ophthalmoscope and the Direct Ophthalmoscope. J Pediatr Ophthalmol Strabismus. 2018;55(3):201–206. [PubMed: 29796680]

39. Saha SK, Fernando B, Cuadros J, Xiao D, Kanagasingam Y. Automated Quality Assessment of Color Fundus Images for Diabetic Retinopathy Screening in Telemedicine. J Digit Imaging. In press.

A deep convolutional neural network can quickly and accurately assess the quality of retinal fundus images acquired during routine retinopathy of prematurity screenings. This may significantly impact telemedicine and computer-based image diagnosis.

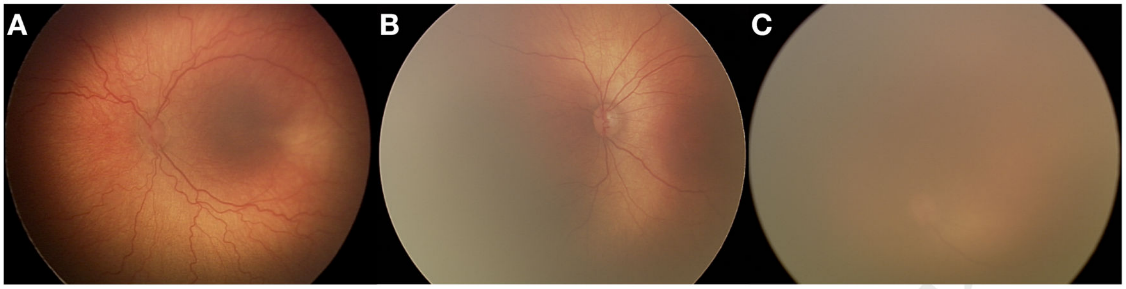


Figure 1: Varying qualities of retinal fundus images.

Representative images from the (A) Acceptable Quality (AQ), (B) Possibly Acceptable Quality (PAQ), and (C) Not Acceptable Quality (NAQ) classes. Note that as image quality degrades, visualization of retinal vasculature becomes more complex, if not impossible. Because NAQ images were not highly represented in our data set (5.6%), they were grouped with the PAQ images into a single category. The final representation of AQ and PAQ images in our data set was 50.8% and 49.2%, respectively.

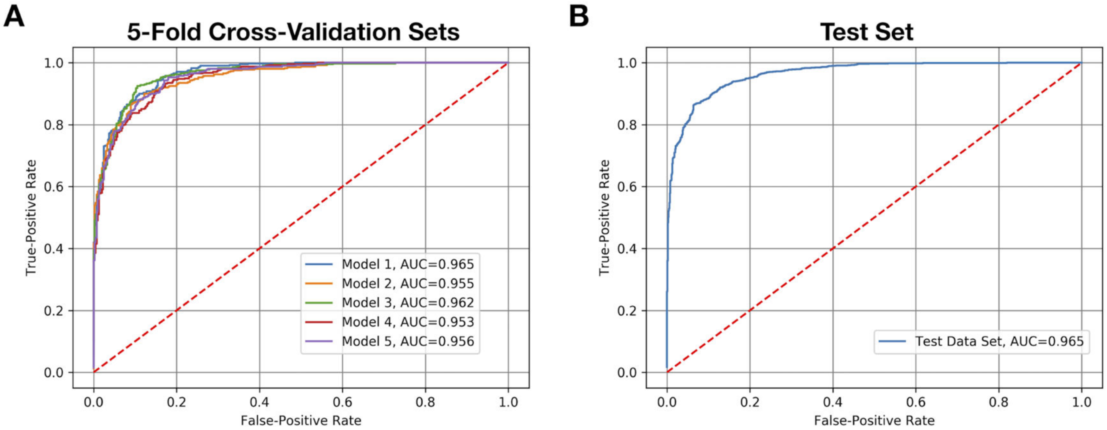
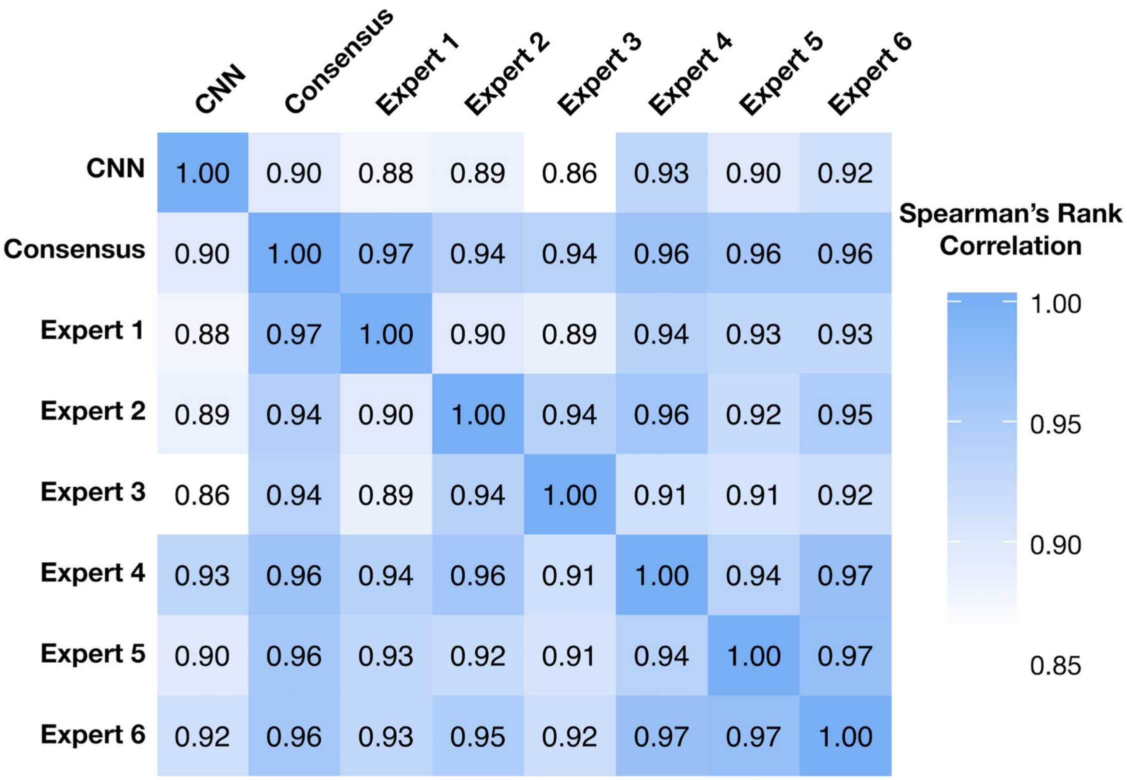


Figure 2: Areas under the receiver operating characteristics curves (AUC).

(A) The AUCs for each convolutional neural network (CNN) produced by 5-fold cross-validation are shown, with mean (SD) equal to 0.958 (0.005). Model 1 demonstrated the highest level of discriminatory power between acceptable quality images and possibly acceptable quality images, as was indicated by the AUC. Therefore, it was selected for final evaluation on the independent test set (B), where it performed with an AUC equal to 0.965, a sensitivity of 93.9% and a specificity of 83.6%.



	CNN	Consensus	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Expert 6
CNN	1.00	0.90	0.88	0.89	0.86	0.93	0.90	0.92
Consensus	0.90	1.00	0.97	0.94	0.94	0.96	0.96	0.96
Expert 1	0.88	0.97	1.00	0.90	0.89	0.94	0.93	0.93
Expert 2	0.89	0.94	0.90	1.00	0.94	0.96	0.92	0.95
Expert 3	0.86	0.94	0.89	0.94	1.00	0.91	0.91	0.92
Expert 4	0.93	0.96	0.94	0.96	0.91	1.00	0.94	0.97
Expert 5	0.90	0.96	0.93	0.92	0.91	0.94	1.00	0.97
Expert 6	0.92	0.96	0.93	0.95	0.92	0.97	0.97	1.00

Figure 3. Correlation heatmap of expert image rankings versus the convolutional neural network (CNN).

The correlation matrix shows Spearman's correlation coefficient values between the CNN image ranking, individual expert grader's image ranking, and the expert graders' consensus ranking. Experts were highly correlated with one another and the overall consensus ranking. The CNN performed nearly as well as individual experts on the ranked data set, as is demonstrated by the high correlation value to the expert consensus ranking.