Supplementary Online Content

Burlina PM, Joshi N, Pacheco KD, Freund DE, Kong J, Bressler NM. Use of Deep Learning for Detailed Severity Characterization and Estimation of 5-Year Risk Among Patients With Age-Related Macular Degeneration. *JAMA Ophthalmol.* Published online September 14, 2018. doi:10.1001/jamaophthalmol.2018.4118

This supplementary material has been provided by the authors to give readers additional information about their work.

eMethods. Additional Points of Clarification

Additional points of clarification not contained in the manuscript are included below:

Details on Network Architecture and Training Methodology

For deep learning, we use the Keras framework (https://keras.io) and make use of its native ResNet50 model. We additionally utilize the native weights of the network, which were pre-trained on the ImageNet dataset, and we then fine-tune these weights with retinal images for the tasks that are reported in the paper, a process referred to as transfer learning. While ImageNet images differ substantially from retinal fundus images, the use of pre-trained weights still helps, most likely in early layers of the network, which focus on basic edge detection like filters, rather than understanding the semantics of the image. We replace the last softmax layer of the network with one that does softmax but contains a number of outputs equal to the number of classes in our experiments (or, alternatively, only a single-node linear activation layer for the regression methods).

For model training, we use a stochastic gradient descent (SGD) optimizer, with Nesterov momentum of 0.9, to minimize a categorical cross entropy loss function (or a mean squared error loss function for the regression case). During training, we utilize a dynamic learning rate schedule, which multiplies the learning rate by 0.5 when the training loss doesn't improve for 10 epochs. Additionally, if the validation loss does not improve for 20 epochs, training is stopped, and the model snapshot (saved on every epoch) with the best validation loss (the dataset was previously subdivided into train/validation/test) is saved as the final model weights. Through hyperparameter optimization on our validation set, we settled on a base learning rate of 0.001 and a batch size of 32.

We additionally employ data augmentation on each batch fed into the network during training. Images are augmented with horizontal flipping, small amounts of blurring or sharpening, as well as various adjustments to saturation, brightness, contrast, and color balance.

Details on Retinal Pre-processing

We used a method that is similar to that reported in[17]. in which we used a preprocessing of the input fundus image by detecting the outer boundaries of the retina, cropping images to the square that is inscribed within the retinal boundary, and resizing the square to fit the expected input size of the network. Here the expected input size for the ResNet network is 224 x 224 pixels[20].

Classification with Additional Classes

The advanced form of AMD is considered level 10 or higher. The main focus of this study is on classes 1-9 because our stated end goal is to predict the probability for the eye to advance to levels 10-11-12. If an eye is already a 10-11-12, then the probability that it advances to the severe stages of 10-11-12 is trivially equal to 1. In the interest of completeness, we succinctly note that the methods herein can also apply to the 12-class problem, and yield (improved) metrics as follows: accuracy (CI error margin): 60.12% (1.17) and linear weighted kappa = 0.8057.

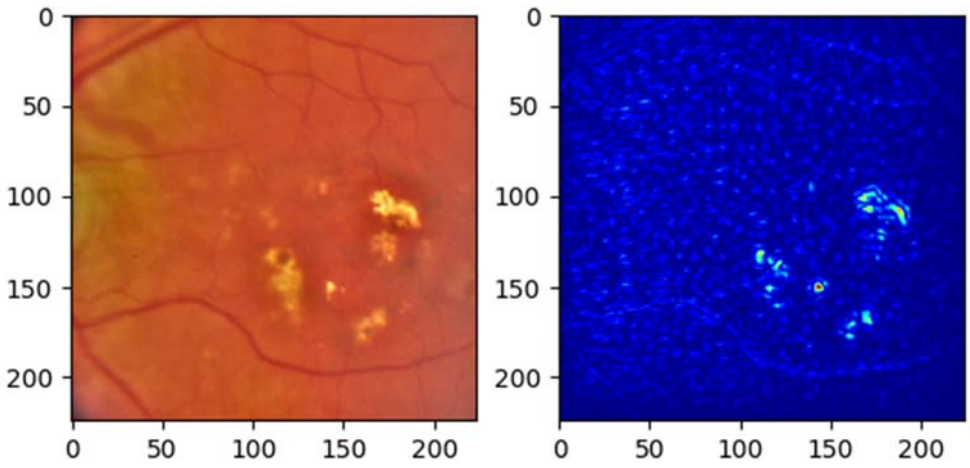
By comparison to another recently published investigation[18] (not available at the time of this submission) of DCCNs applied to a 12+1-step AREDS detailed scale (where the additional classes correspond to steps 10,11 and 12, as well as a class for ungraded images), our study focusses on the novel task of predicting the 5-year risk probability via the 9-step scale. Both studies however point to somewhat related levels of classification performance considering various factors mentioned in the discussion section.

Potential Deployment Scenarios

Regarding real world clinical deployment and what classification should be used, several possibilities can be envisaged, including for example: having a single 12 class classifier as noted earlier; having a 13 class classifier with a 'bad quality' class; having a cascaded decision process with several classifiers: one that checks quality and recommends that the fundus imager retake an image in case quality is insufficient, followed by a second stage classifier that predicts the current state of the eye based on a 4 class classification (as we report in this paper) that would help refer individuals to be seen or not by a specialist, this would then be followed by another stage predictor that would work in the clinic and predict the probability of advancing if the eye is not already in the advanced stage (via the 9 class predictor we address in this study).The first and second stages could work automatically in a kiosk, teleophthalmology, point of care, or self scanning scenario, and the third stage would work in the clinic, and help a clinician decide on a course of action. In the end, how one would actually structure classification for deployment really would depend on the actual clinical workflow. Our inclination is that a three stage system, with a human clinician in the loop and a machine assisting him/her is probably a potentially viable approach for deployment. But many other choices are possible.

Saliency Maps

The problem of interpretably of DCNNs and addressing their inherent black box-ness is important to ensure trust in their workings and allow their future use in clinical applications. Research on tools that allow for greater interpretation of DCNNs workings is still ongoing. However, saliency maps can offer a simple way to visualize what the network activates on. As the example shown in the eFigure 1 suggests, the saliency map points to the network likely activating on lesions and specific drusen, which seems therefore to indicate that the network is probably looking at the "right'' locations when making a decision.



Supplemental Figure 1. Fundus image (left) and corresponding saliency map (right).

Kappa weightings

We used linear weighted kappa, which is commonly used[23]. This provides a fair assessment when the classes are ordinal when compared to the more pessimistic unweighted version of kappa, and the optimistic quadratically weighted version of kappa. For comparison, we provide here these additional figures: for the 9 step problem the un-weighted kappa was 0.4457; The linear weighted kappa was 0.7382; The quadratic weighted kappa was 0.8657.

Human Annotations

Human performance evaluation: The human annotated images and test images used for the algorithm for the 4 class problem in Figure 1 are sampled randomly from AREDS exactly as done and reported in reference[17], and as such, while statistically "comparable'', the comparison is not done on the exact same set of images.

Visits during AREDS

For each visit, a stereoscopic pair of images was taken for each eye and, for each stereoscopic pair, we only used one of the stereo images per eye. Also note: In some visits only one eye was imaged in the AREDS study. There are a variable number of visits for each subject, some individuals attended for the full 12 years, while some were much shorter lived.

AREDS Analog Image Conversion

Note that the AREDS images used here were originally taken from analog images that were subsequently digitally scanned. Performance on digital images may therefore vary and may entail improvement or worsening, but likely would lead to improved performance. More discussions on this topic can be found in: Automated Classification of Severity of Age-Related Macular Degeneration from Fundus Photographs, by Kankanahalli, Srihari, Burlina, Philippe, Wolfson, Yulia; et al. IOVS, Vol. 54, Issue: 3, pp 1789-1796, 2013

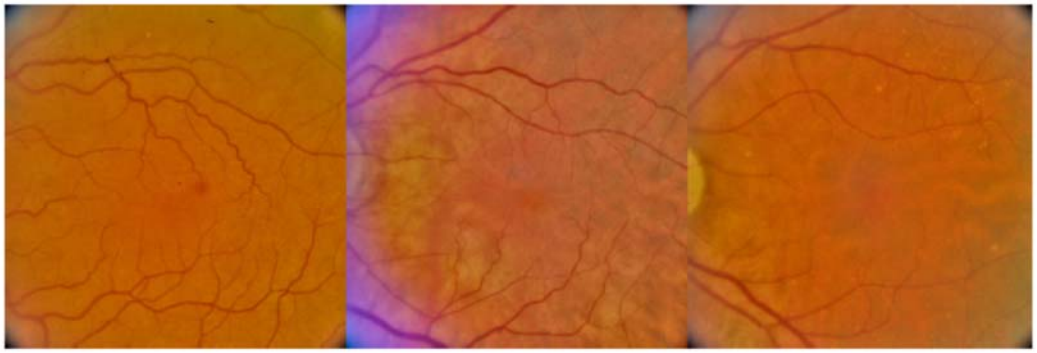
Model weights

Model weights may be shared with interested parties if used solely for research and non-commercial and non-clinical/diagnostic purposes and provided that appropriate usage and legal safeguards are employed. Request should be made by contacting the corresponding author.

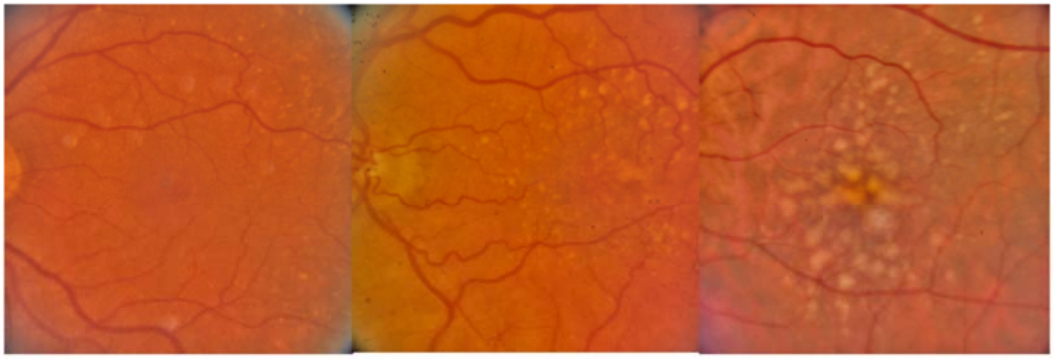
eTable. The 9-step AMD severity scale from AREDS report 17[5]

Step	Total Drusen Area	Increased Pigment	Depigmentation-GA	5-year risk factor (%)
1	<C-1	0	0	0.3
2	≥C-1; <C-2 <C-1	0 ≥Q	0 ≥Q; <I-2	0.6
3	≥C-2; <I-2	0	0	1.9
4	≥I-2; <O-2 ≥C-1; <I-2 < C-2	0 ≥Q ≥0	0 ≥Q; <I-2 ≥I-2; <0.5DA	4.9
5	≥O-2; <0.5DA ≥I-2; <O-2 ≥C-2; <I-2	0 ≥Q ≥0	0 ≥Q; <I-2 ≥I-2; <0.5DA	6.1
6	≥0.5DA ≥O-2; <0.5DA ≥I-2; <O-2	0 ≥Q ≥0	0 ≥Q; <I-2 ≥I-2; <0.5DA	13.9
7	≥0.5DA ≥O-2; <0.5DA	≥Q ≥0	≥Q; <I-2 ≥I-2; <0.5DA	28.1
8	≥0.5DA Any	≥0 ≥0	≥I-2; <0.5DA ≥0.5DA	47.4
9	Any	≥0	Noncentral GA	53.2

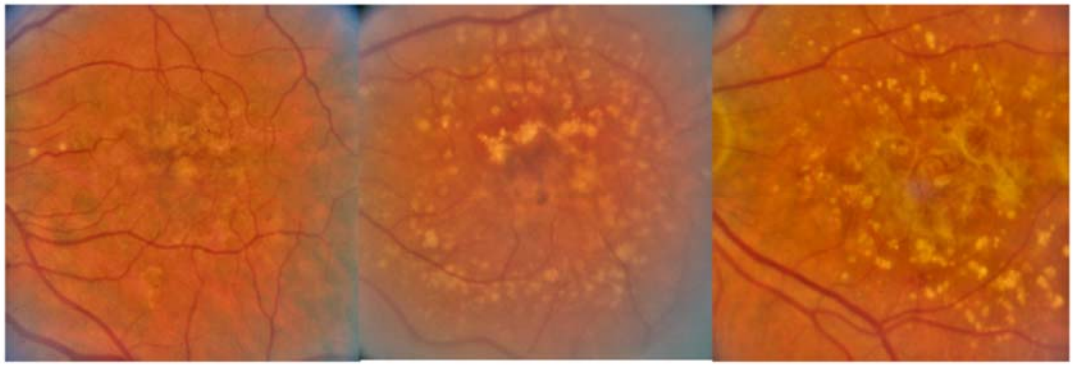
Generic Abbreviations: DA=disc area; GA=geographic atrophy; Q=questionable category which means grader is at least 50%, but not 90%, sure abnormality exits;
Abbreviations for standard circles used for measuring size and area of abnormalities:
C-1 125μm and 0.0069DA; C-2 250μm and 0.028DA; I-2 354μm and 0.056DA; O-2 650μm and 0.19DA; 0.5DA 1061μm and 0.5DA.



eFigure 1. Examples of fundus images showing age-related macular degeneration (AMD) 9-step classification ranging from 1 through 3 (left-right).



eFigure 2. Examples of fundus images showing age-related macular degeneration (AMD) 9-step classification ranging from 4 through 6 (left-right).



eFigure 3. Examples of fundus images showing age-related macular degeneration (AMD) 9-step classification ranging from 7 through 9 (left-right).