


# Automated Grading of Age-Related Macular Degeneration From Color Fundus Images Using Deep Convolutional Neural Networks

Philippe M. Burlina, PhD; Neil Joshi, BS; Michael Pekala, MS; Katia D. Pacheco, MD; David E. Freund, PhD; Neil M. Bressler, MD

 CME Quiz at [jamanetwork.com/learning](http://jamanetwork.com/learning)

**IMPORTANCE** Age-related macular degeneration (AMD) affects millions of people throughout the world. The intermediate stage may go undetected, as it typically is asymptomatic. However, the preferred practice patterns for AMD recommend identifying individuals with this stage of the disease to educate how to monitor for the early detection of the choroidal neovascular stage before substantial vision loss has occurred and to consider dietary supplements that might reduce the risk of the disease progressing from the intermediate to the advanced stage. Identification, though, can be time-intensive and requires expertly trained individuals.

**OBJECTIVE** To develop methods for automatically detecting AMD from fundus images using a novel application of deep learning methods to the automated assessment of these images and to leverage artificial intelligence advances.

**DESIGN, SETTING, AND PARTICIPANTS** Deep convolutional neural networks that are explicitly trained for performing automated AMD grading were compared with an alternate deep learning method that used transfer learning and universal features and with a trained clinical grader. Age-related macular degeneration automated detection was applied to a 2-class classification problem in which the task was to distinguish the disease-free/early stages from the referable intermediate/advanced stages. Using several experiments that entailed different data partitioning, the performance of the machine algorithms and human graders in evaluating more than 130 000 images that were deidentified with respect to age, sex, and race/ethnicity from 4613 patients against a gold standard included in the National Institutes of Health Age-Related Eye Disease Study data set was evaluated.

**MAIN OUTCOMES AND MEASURES** Accuracy, receiver operating characteristics and area under the curve, and  $\kappa$  score.

**RESULTS** The deep convolutional neural network method yielded accuracy that ranged between 88.4% (SD, 0.5%) and 91.6% (SD, 0.1%), the area under the receiver operating characteristic curve was between 0.94 and 0.96, and  $\kappa$  (SD) between 0.764 (0.010) and 0.829 (0.003), which indicated a substantial agreement with the gold standard Age-Related Eye Disease Study data set.

**CONCLUSIONS AND RELEVANCE** Applying a deep learning-based automated assessment of AMD from fundus images can produce results that are similar to human performance levels. This study demonstrates that automated algorithms could play a role that is independent of expert human graders in the current management of AMD and could address the costs of screening or monitoring, access to health care, and the assessment of novel treatments that address the development or progression of AMD.

JAMA Ophthalmol. 2017;135(11):1170-1176. doi:10.1001/jamaophthalmol.2017.3782  
Published online September 28, 2017.

**Author Affiliations:** The Johns Hopkins University Applied Physics Laboratory, Laurel, Maryland (Burlina, Joshi, Pekala, Freund); Retina Division, Brazilian Center of Vision Eye Hospital, Brasília, DF, Brazil (Pacheco); Retina Division, Wilmer Eye Institute, Johns Hopkins University School of Medicine, Baltimore, Maryland (Bressler); Editor, *JAMA Ophthalmology* (Bressler).

**Corresponding Author:** Neil M. Bressler, MD, Wilmer Eye Institute, Johns Hopkins University, 600 N Wolfe St, Maumenee 752, Baltimore, MD 21287-9227 ([nmboffice@jhmi.edu](mailto:nmboffice@jhmi.edu)).

Age-related macular degeneration (AMD) is associated with the presence of drusen, long-spacing collagen, and phospholipid vesicles between the basement membrane of the retinal pigment epithelium and the remainder of the Bruch membrane.<sup>1</sup> The intermediate stage of AMD, which often causes no visual deficit, includes eyes with many medium-sized drusen (the greatest linear dimension ranging from 63  $\mu\text{m}$ -125  $\mu\text{m}$ ) or at least 1 large druse (greater than 125  $\mu\text{m}$ ) or geographic atrophy (GA) of the retinal pigment epithelium that does not involve the fovea.<sup>1</sup>

The intermediate stage often leads to the advanced stage, in which substantial damage to the macula can occur from choroidal neovascularization, also termed the *wet* advanced form, or GA that involve the center of the macula, which is termed the *dry* advanced form. Choroidal neovascularization, when not treated, often leads to the loss of central visual acuity,<sup>2</sup> which affects daily activities like reading, driving, or recognizing objects. Consequently, the advanced stage can pose a substantial socioeconomic burden on society.<sup>3</sup> Age-related macular degeneration is the leading cause of central vision loss among people older than 50 years in the United States; approximately 1.75 million to 3 million individuals have the advanced stage.<sup>3-5</sup>

While AMD currently has no definite cure, the Age-Related Eye Disease Study (AREDS) has suggested benefits of specific dietary supplements for slowing AMD progression among individuals with the intermediate stage in at least 1 eye or the advanced stage only in 1 eye.<sup>6</sup> Additionally, vision loss because of choroidal neovascularization can be reversed, stopped, or slowed by administering antivascular endothelial growth factor intravitreal injections.<sup>7</sup> Ideally, individuals with the intermediate stage of AMD should be identified, even if asymptomatic, and referred to an ophthalmologist who can monitor for the development and subsequent treatment of choroidal neovascularization. Manual screenings of the entire at-risk population of individuals older than 50 years for the development of the intermediate stage of AMD in the United States is not realistic because the at-risk population is large (more than 110 million).<sup>8</sup> It also is not feasible in all US health care environments to screen if there is poor access to experts who can identify the development of the intermediate stage of AMD. These same issues may be more pronounced in low- and middle-income countries. Therefore, automated AMD diagnostic algorithms, which identify the intermediate stage of AMD, are a worthy goal for future automated screening solutions for major eye diseases.

While no treatment comparable with antivascular endothelial growth factor currently exists for GA, numerous clinical trials are being conducted to identify treatments for slowing GA growth.<sup>9-12</sup> Automated algorithms may play a role in assessing treatment efficacy, in which it is critical to quantify disease worsening objectively under therapy; careful manual grading of this by clinicians can be costly and subjective.

Past algorithms for automated retinal image analysis generally relied on traditional approaches that consisted of manually selecting engineered image features (eg, wavelets, scale-invariant feature transform<sup>13-15</sup>) that were then used in a classifier<sup>13-20</sup> (eg, support vector machines [SVM]<sup>15,16</sup> or ran-

## Key Points

**Question** When applying deep learning methods to the automated assessment of fundus images, what is the accuracy for detecting age-related macular degeneration?

**Finding** This study found that the deep convolutional neural network method ranged in accuracy between 88.4% (SD, 0.7%) and 91.6% (SD, 0.1%), with  $\kappa$  scores close to or greater than 0.8, which is comparable with human expert performance levels.

**Meaning** The results suggest that deep learning-based machine grading can be leveraged successfully to automatically assess age-related macular degeneration from fundus images in a way that is comparable with the human ability to grade age-related macular degeneration from these images.

dom forests<sup>14</sup>). By contrast, deep learning (DL) methods<sup>17,21-29</sup> learn task-specific image features with multiple levels of abstraction without relying on manual feature selection. Recent advances in DL have improved performance levels dramatically for numerous image analysis tasks. This progress was enabled by many factors (eg, novel methods to train very deep networks or using graphic processing units).<sup>22-26</sup> Recently, DL has been used for conducting retinal image analyses, including tasks such as classifying referable diabetic retinopathy.<sup>27,28</sup> A previous study<sup>17,21</sup> reported on the use of deep universal features/transfer learning for automated AMD grading. The new study expanded on the previous study by using a data set that is approximately 10 to 20 times larger, using the full scope of deep convolutional neural networks (DCNN).

## Methods

### Overview

This study aimed to solve a 2-class AMD classification problem, classifying fundus images of individuals that have either no or early stage AMD (for which dietary supplements and monitoring for progression to advanced AMD is not considered) vs those with the intermediate or advanced stage AMD, for which supplements, monitoring, or both is considered. It leveraged DL and DCNN. The goals of this study were to measure and compare the performance of the proposed DL vs a human clinician, and a secondary goal was to compare the performance between 2 DL approaches that entailed different levels of computational effort regarding training.

### Data

Our study used the National Institutes of Health AREDS data set collected over a 12-year period. AREDS originally was designed to improve understanding of AMD worsening, treatment, and risk factors for worsening. It includes more than 130 000 color fundus images from 4613 patients that were taken with written informed consent obtained at each of the clinical sites (Table 1). Color fundus photographs were captured of each patient at baseline and follow-up visits and were subsequently digitized. These images included stereo pairs

Table 1. Summary of Data Sets Used

Data Set	H <sup>a</sup>	WS <sup>b</sup>	NSG <sup>c</sup>	NS <sup>d</sup>
No. of images				
Class 0	2779	74 401	37 101	37 418
Class 1	2221	59 420	29 842	29 983

Abbreviations: AREDS, Age-Related Eye Disease Study; DCNN, deep convolutional neural networks; H, human; NS, no stereo; NSG, no stereo gradable; WS, with stereo pairs.

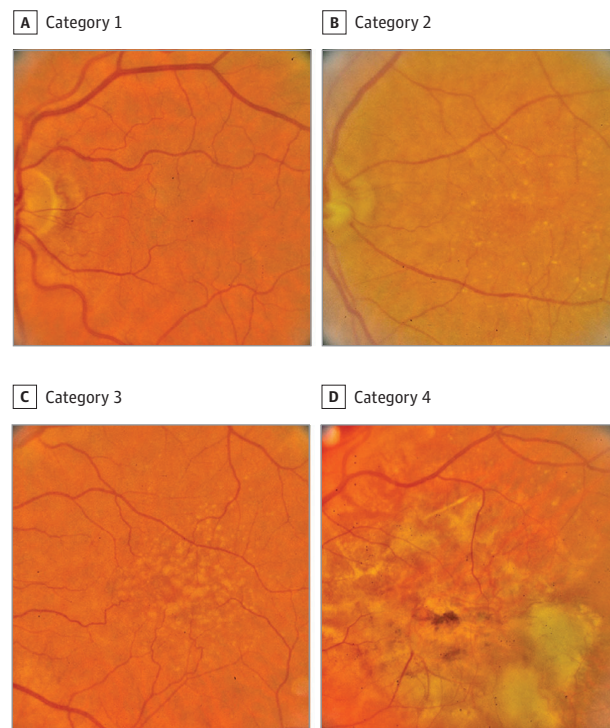
<sup>a</sup> For comparing DCNN algorithms with human performance, a physician independently and manually graded a subset ( $n = 5000$ ) of the AREDS images.

<sup>b</sup> This is the full set of all AREDS images ( $n = 133\,821$ ) including stereo pairs (taking care that stereo pair images from the same eye did not appear in the training and testing data sets).

<sup>c</sup> Because AREDS images are collected under a variety of environmental conditions (eg, lighting, patient eye orientation, etc) and therefore are not of uniform quality, an ophthalmologist was tasked to annotate a subset ( $n = 7775$ ) of images for “gradability” as a basic measure of fundus image quality. This metric was extended via machine learning over the entire data set and 458 of the poorest-quality images were removed from NS to form NSG.

<sup>d</sup> Only 1 of the stereo pair is kept from each eye resulting in a set comprising 67 401 images.

Figure 1. Examples of Fundus Images Showing Age-Related Macular Degeneration (AMD).



A, Category 1 or no AMD; B, category 2 or early AMD; C, category 3, intermediate AMD; and D, category 4 or advanced AMD.

taken from both eyes. Images were carefully and quantitatively graded by experts for identifying AMD at a US fundus photograph-reading center.<sup>2</sup> Graders used graduated circles to measure the location and area of drusen and other retinal abnormalities (eg, retinal elevation and pigment abnormalities) in the fundus images to determine the AMD severity level.<sup>2</sup> Each image was then assigned by graders to a category reflecting AMD severity that ranged from 1 to 4, with 1 = no AMD, 2 = early stage, 3 = intermediate stage, and 4 = advanced stage (Figure 1). These severity grades were used as a “gold standard” in our study for performing a 2-class classification of no or early stage AMD (here referred to as class 0) vs potentially

referable (intermediate or advanced) stage (class 1). AREDS is a public data set that can be made available on request to the National Institutes of Health.

### DCNN Approach

This study used DCNNs. A DCNN is a deep neural network that consists of many repeated processing layers that take as input fundus images that are processed via a cascade of operations with the goal of producing an output class label for each image.<sup>23,25,26</sup> One way to think about DCNNs is that they match the input image with successive convolutional filters to generate low-, mid-, and high-level representations (ie, features) of the input image. Deep convolutional neural networks also include layers that pool features together spatially, perform nonlinear operations at various levels, combine these via fully connected layers, and output a final probability value for the class label (here the AMD-referable vs not referable classification). A DCNN is trained to discover and optimize the weights of the convolutional filters that produce these image features via a backpropagation process. This optimization is done directly by using the training images. Therefore, this process is considered to be a data-driven approach and contrasts with past approaches to processing and analyzing fundus imagery that have used engineered features that resulted from an ad hoc, manual, and therefore possibly suboptimal algorithmic design and selection of such features. While the workings of DCNNs are simple to grasp at a notional level, there is currently extensive research being conducted to understand, improve, and extend the current state of the art.

We used the AlexNet (University of Toronto) DCNN model (here called DCNN-A)<sup>23</sup> in which the weights of all layers of the network are optimized via training to solve the referable AMD classification problem. This training process involved optimizing more than 61 million convolutional filter weights. In addition to the layers mentioned above, this network included dropout, rectified linear unit activation, and contrast normalization steps.<sup>23</sup> The dropout step consisted of arbitrarily setting to 0 some of the neuron outputs (chosen randomly) with the effect of encouraging functional redundancy in the network and acting as a regularization. Our implementation incorporated the Keras and TensorFlow DL frameworks. It used a stochastic gradient descent with a Nesterov momentum, with an initial learning rate that was set to

0.001. The training scheme used an early stopping mechanism that terminated training after 50 epochs of no improvement of the validation accuracy.<sup>23</sup>

### Universal Features/Transfer Learning Approach

For comparison, this study also used another DL approach that focused on reusing a pretrained DCNN and performing transfer learning.<sup>21,30</sup> The idea behind transfer learning is to exploit knowledge that is learned from one source task that has a relative abundance of training data (general images of animals, food, etc.) to allow for learning in an alternative target task (AMD classification on fundus images). Here, universal features were computed by using a pretrained DCNN to solve a general classification problem on a large set of images and reuse these features for the AMD task. Our approach<sup>17,21</sup> used the pretrained OverFeat (New York University)<sup>24</sup> DCNN, which was pretrained on more than a million natural images to produce a 4096 dimension feature vector, which was then used to retrain a linear SVM (LSVM)<sup>17,21,24</sup> for our specific AMD classification problem from fundus images. We call this method DCNN-U.

The 2 methods (DCNN-A and DCNN-U) used a preprocessing of the input fundus image by detecting the outer boundaries of the retina, cropping images to the square that was inscribed within the retinal boundary, and resizing the square to fit the expected input size of AlexNet or OverFeat DCNNs. Additionally, DCNN-U used a multigrid approach in which the cropped image was coupled with 2 concentric square subimages that were centered in the middle of the inscribed image. The resulting 3 images (the cropped image plus 2 centered subimages) were then fed to the OverFeat DCNN to produce 2 additional 4096-long feature vectors. The 3 feature vectors for the image were then concatenated to generate a single 12 288-sized feature vector as input to the LSVM. This method is further detailed in previous reports.<sup>17,21</sup>

### Data Partitioning

This study considered several experiments that used the entire AREDS fundus image data set as well as different subsets of AREDS. It also used different partitionings and groupings of the AREDS image data set. The different subsets of AREDS used are described here. The set of all AREDS images (133 821) was used, including stereo pairs (ensuring that stereo pairs from the same eye did not appear in the training and testing data sets). We called this set WS for “with stereo pairs.” We called the next set NS for “no stereo.” In this data set, only 1 of the stereo images was kept from each eye, which resulted in 67 401 images. We called the next set NSG for “no stereo, gradable.” Because AREDS images are collected under a variety of conditions (eg, lighting or eye orientation) and therefore are not of uniform quality, an ophthalmologist (K.D.P.) was tasked to annotate a subset of images ( $n = 7775$ , 5.8%) for “gradability” as a basic measure of fundus image quality. Subsequently, a machine learning method was used to extend the index of gradability over the entire image data set NS to exclude automatically the most egregious low-quality images. The NSG was derived from NS by removing 458 images (0.34%) with the smallest “gradability” index. The final set was called H for hu-

man. For comparison with human performance levels, we tasked a physician to independently and manually grade a subset of AREDS images ( $n = 5000$ , 3.7%). The grades that were generated by the physician and the machine were compared with the AREDS gold standard AMD scores. The number of images that were used in each set, broken down by class, is reported in Table 1.

These data sets were further subdivided into training and testing subsets. We used a conventional K-fold crossvalidation performance evaluation method, with  $K = 5$ , in which 4 folds were used for training and 1 was used for testing (with a rotation of the folds). Additionally, because images from patients were collected over multiple visits, and because DCNN performance depends on having as large a number as possible of patient examples, we considered 2 types of experiments that corresponded to 2 types of data grouping and partitioning. In the baseline partitioning method (termed *standard partitioning* [SP]) images taken at each patient visit (occurring approximately every 2 years) were considered unique. For SP, when both stereo pairs were used (WS), care was taken that they always appeared together in the same fold. In a second partitioning method (termed *patient partitioning* [PP]), we ensured that all images of the same patient appeared in the same fold. Standard partitioning views patient visit as a unique entity, while PP considers that each patient (not each visit) forms a unique entity. Therefore, PP is a more stringent partitioning method that provides fewer patients to the classifier to train on; any patient with a highly abnormal or atypical retina will be represented in only 1 of the folds.

### Performance Metrics

The performance metrics used included accuracy, sensitivity, specificity, positive predicted value, negative predicted value, and  $\kappa$  score, which accounts for the possibility of agreement by chance.<sup>1,31</sup> Because any classifier trades off between sensitivity and specificity, to compare methods we used receiver operating characteristic (ROC) curves that plot the detection probability, ie, sensitivity vs false alarm rate (ie, 100% minus specificity) for each algorithm/experiment. To compare with human performance levels, we also showed the operating point that demonstrated the human clinician operating performance level. We also computed the area under the curve for each algorithm/experiment.

## Results

The experiments used the AREDS fundus images with the different subsets and partitioning that were previously explained. Performance levels are reported in Table 2 (SP) and Table 3 (PP) for sets H, WS, NS, and NSG, and for the 2 algorithms (DCNN-A and DCNN-U) and the human performance levels. Receiver operating characteristic curves and areas under the curve are reported in Figure 2.

In aggregate, performance results for both DL approaches show promising outcomes when considering all metrics. Accuracy ranged from 90.0% (SD, 0.6%) to 91.6% (SD, 0.1%) for DCNN-A (Table 2) and 88.4% (SD, 0.5%) to 88.8% (SD, 0.7%)



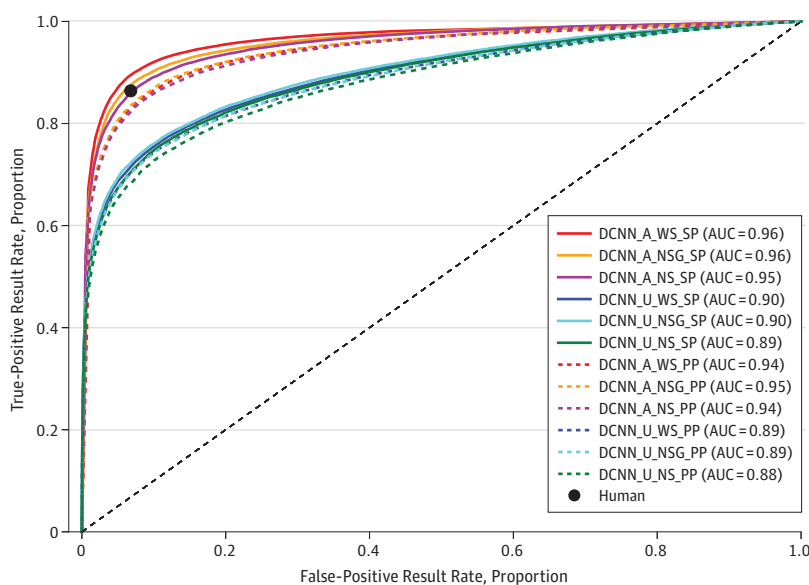
Table 2. Performance Levels for Human and Machine Experiments Using Standard Partitioned Data<sup>a</sup>

Method/Data Set	Human H	DCNN-A WS	DCNN-U WS	DCNN-A NSG	DCNN-U NSG	DCNN-A NS	DCNN-U NS
Accuracy	90.2	91.6 (0.1)	83.7 (0.5)	90.7 (0.5)	83.9 (0.4)	90.0 (0.6)	83.2 (0.2)
Sensitivity	86.4	88.4 (0.7)	73.5 (0.9)	87.2 (0.8)	73.8 (0.7)	85.7 (2.3)	72.8 (0.2)
Specificity	93.2	94.1 (0.6)	91.8 (0.3)	93.4 (1.0)	92.1 (0.5)	93.4 (1.0)	91.5 (0.2)
PPV	91.0	92.3 (0.7)	87.7 (0.2)	91.5 (1.2)	88.3 (0.7)	91.3 (1.0)	87.3 (0.3)
NPV	89.6	91.1 (0.4)	81.2 (0.7)	90.1 (0.4)	81.4 (0.4)	89.1 (1.4)	80.8 (0.1)
κ	0.800	0.829 (0.003)	0.663 (0.010)	0.810 (0.011)	0.700 (0.008)	0.796 (0.013)	0.654 (0.003)

Abbreviations: DCNN-A, deep convolutional neural network, algorithm A; DCNN-U, deep convolutional neural network, algorithm U; H, human; NS, no stereo; NSG, no stereo gradable; NPV, negative predicted value; PPV, positive predicted value; WS, with stereo pairs.

<sup>a</sup> All values indicate percentages, except for κ. Values in parentheses indicate standard deviations.

Figure 2. Receiver Operating Characteristic Curves



Receiver operating characteristic curves for all experiments and algorithms showing also the corresponding area under the curve values. A indicates algorithm A; AUC, area under the curve; DCNN, deep convolutional neural networks; NS, no stereo; NSG, no stereo gradable; PP, patient partitioning; SP, standard partitioning; WS, with stereo pairs; U, algorithm U.

Table 3. Performance Levels for Human and Machine Experiments Using Patient Partitioned Data<sup>a</sup>

Method/Data Set	Human H	DCNN-A WS	DCNN-U WS	DCNN-A NSG	DCNN-U NSG	DCNN-A NS	DCNN-U NS
Accuracy	90.2	88.7 (0.7)	83.1 (0.9)	88.8 (0.7)	83.1 (0.5)	88.4 (0.5)	82.4 (0.5)
Sensitivity	86.4	84.6 (0.9)	72.3 (2.2)	85.3 (1.6)	71.7 (1.4)	84.5 (0.9)	71.0 (1.3)
Specificity	93.2	92.0 (0.7)	91.8 (0.6)	91.6 (1.2)	92.2 (0.5)	91.5 (0.7)	91.4 (0.3)
PPV	91.0	89.4 (1.1)	87.5 (1.1)	89.2 (1.1)	88.0 (0.7)	88.9 (1.0)	86.9 (0.5)
NPV	89.6	88.2 (1.0)	80.6 (1.4)	88.6 (1.1)	80.2 (1.1)	88.0 (0.5)	79.8 (0.5)
Kappa	0.800	0.770 (0.013)	0.652 (0.020)	0.773 (0.014)	0.651 (0.010)	0.764 (0.010)	0.636 (0.011)

Abbreviations: DCNN-A, deep convolutional neural network, algorithm A; DCNN-U, deep convolutional neural network, algorithm U; H, human; NS, no stereo; NSG, no stereo gradable; NPV, negative predicted value; PPV, positive predicted value; WS, with stereo pairs.

<sup>a</sup> All values indicate percentages, except for the κ. Values in parentheses indicate standard deviations.

(Table 3); for DCNN-U, it ranged from 83.2% (SD, 0.2%) to 83.9% (SD, 0.4%) (Table 2) and 82.4% (SD, 0.5%) to 83.1% (SD, 0.5%) (Table 3). As seen in Table 2, Table 3, and the ROCs, DCNN-A consistently outperformed DCNN-U. This can be explained by the fact that DCNN-A was specifically trained to solve the AMD classification problem by optimizing all of the DCNN weights

over all layers of the network, while for DCNN-U, with its simpler training requirement, the training only affected the final (LSVM) classification stage.

Table 2 and Table 3 also suggest that the DCNN-A results are comparable with human performance levels. Based on accuracy and κ scores, in Table 3, DCNN-A performance (accu-

racy = 88.7% [0.7];  $\kappa$  = 0.770 [0.013]) is close or comparable with human performance levels (accuracy = 90.2% and  $\kappa$  = 0.800), and in Table 2 it exceeds slightly the human performance levels (accuracy = 91.6% [0.1],  $\kappa$  = 0.829 [0.003]). In Table 2 and 3, the  $\kappa$  scores for DCNN-A ( $\kappa$  = 0.764 [0.010]-0.829 [0.003]) and the human grader ( $\kappa$  = 0.800) show substantial to near perfect agreement with the AREDS AMD gold standard grading, while DCNN-U exhibits substantial agreement ( $\kappa$  = 0.636 [0.011]-0.700 [0.008]). Receiver operating characteristic curves also show similar human and machine performance levels. The other metrics in Table 2 and Table 3 also echo these observations.

To test algorithms on images that are representative of the quality that one would expect in actual practice, we did not perform extensive eliminations of images based on their quality. In particular, data sets WS and NS used all images while data set NSG removed only 458 (approximately 0.68%) of the worst-quality images. When looking at the performance of NSG vs NS, there was a small but measurable decrease in performance levels, as seen when comparing the accuracy of DCNN-A of 90.7% for NSG vs 90.0% for NS (Table 2).

Experiments that used PP showed a small degradation in performance levels when compared with experiments that used SP. This is because, for patient partitioning, the classifier was trained on 923 fewer patients (20%). The performance in SP was reflective of a scenario in which training would take advantage of knowledge that was gained during a longitudinal study, vs PP experiments that take a strict view on grouping to remove any possible correlation between fundus images across visits. In aggregate, after accounting for network and partition differences, the results that were obtained for WS, NSG, and NS were close, with a preference for WS (since there were more data to train from) and NSG (because some low-quality images were removed) over NS. For example, DCNN-A accuracies are 91.6% (SD, 0.1%) (WS), 90.7% (SD, 0.5%) (NSG), and 90.0% (SD, 0.6%) (NS) (Table 2).

## Discussion

We described using DL methods for the automated assessment of AMD from color fundus images. These experimental results show promising performance levels in which deep convolutional neural networks appear to perform a screening function that has clinical relevance with performance levels that are comparable with physicians. Specifically, the AREDS data set is, to our knowledge, the largest annotated fundus image

data set that is currently available for AMD. Therefore, this study may constitute a useful baseline for future machine-learning methods to be applied to AMD.

## Limitations

One limitation of this data set is a mild class imbalance regarding the number of fundus images in class 1 vs 0, which may have a moderate effect on performance levels. Another potential limitation is that this data set uses digitized images that were taken from analog photographs. This possibly can negatively affect quality and machine performance when compared with digital fundus acquisition, but this possibility cannot be determined from this investigation because none of the images were digital.

Another limitation of this study is that it relies exclusively on AREDS and does not make use of a separately collected clinical data set for performance evaluation, as was done in the diabetic retinopathy studies<sup>27</sup> (eg, training a model on EyePACS [EyePACS LLC] and testing on Methods to Evaluate Segmentation and Indexing Techniques in the Field of Retinal Ophthalmology [MESSIDOR]). The situation is different, however, for AMD in which there is currently no large reference clinical data set for use other than AREDS.

Future clinical translation of DL approaches would require validation on separate clinical data sets and using more human clinicians for comparison. While this study offers a promising foray into using DL for automated AMD analysis, future work could involve using more sophisticated networks to improve performance, expanding to lesion delineation and exploiting other modalities (eg, optical coherence tomography).

## Conclusions

This study showed that automated algorithms can play a role in addressing several clinically relevant challenges in the management of AMD, including cost of screening, access to health care, and the assessment of novel treatments. The results of this study, using more than 130 000 images from AREDS, suggest that new DL algorithms can perform a screening function that has clinical relevance with results similar to human performance levels to help find individuals that likely should be referred to an ophthalmologist in the management of AMD. This approach could be used to distinguish among various retinal pathologies and subsequently classify the severity level within the identified pathology.

## ARTICLE INFORMATION

**Accepted for Publication:** August 1, 2017.

**Published Online:** September 28, 2017.  
doi:10.1001/jamaophthalmol.2017.3782

**Author Contributions:** Dr Burlina and Mr Joshi had full access to all the data in the study and take full responsibility for the integrity of the data and the accuracy of the data analysis.  
*Concept and design:* Burlina, Joshi, Pacheco, Freund, Bressler.  
*Acquisition, analysis, or interpretation of data:*

Burlina, Joshi, Pacheco, Freund, Bressler.

*Drafting of the manuscript:* Burlina, Joshi, Pekala, Pacheco, Freund.

*Critical revision of the manuscript for important intellectual content:* Burlina, Joshi, Freund, Bressler.  
*Statistical analysis:* Burlina, Joshi, Pekala, Pacheco, Freund.

*Obtained funding:* Burlina, Bressler.

*Administrative, technical, or material support:* Burlina, Joshi, Freund.

*Supervision:* Burlina, Pacheco, Bressler.

**Conflict of Interest Disclosures:** All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Drs Burlina, Freund, and Bressler report holding a patent on a system and method for the automated detection of age-related macular degeneration and other retinal abnormalities. No other disclosures were reported.

**Funding/Support:** This work was supported by award R21EY024310 from the National Eye Institute, the James P. Gills Professorship, and

unrestricted research funds to the Johns Hopkins University School of Medicine Retina Division for Macular Degeneration and Related Diseases Research.

**Role of the Funder/Sponsor:** The National Eye Institute and the Johns Hopkins University had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

**Disclaimer:** Dr Bressler is the Editor of *JAMA Ophthalmology*, but he was not involved in any of the decisions regarding review of the manuscript or its acceptance.

**Additional Information:** The AREDS dbGAP dataset was made available from the National Eye Institute of the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## REFERENCES

- Age-Related Eye Disease Study Research Group. The Age-Related Eye Disease Study system for classifying age-related macular degeneration from stereoscopic color fundus photographs: the Age-Related Eye Disease Study Report Number 6. *Am J Ophthalmol*. 2001;132(5):668-681.
- Bird AC, Bressler NM, Bressler SB, et al; The International ARM Epidemiological Study Group. An international classification and grading system for age-related maculopathy and age-related macular degeneration. *Surv Ophthalmol*. 1995;39(5):367-374.
- Bressler NM. Age-related macular degeneration is the leading cause of blindness.... *JAMA*. 2004;291(15):1900-1901.
- Macular Photocoagulation Study Group. Subfoveal neovascular lesions in age-related macular degeneration. guidelines for evaluation and treatment in the macular photocoagulation study. *Arch Ophthalmol*. 1991;109(9):1242-1257.
- Bressler NM, Bressler SB, Congdon NG, et al; Age-Related Eye Disease Study Research Group. Potential public health impact of Age-Related Eye Disease Study results: AREDS report no. 11. *Arch Ophthalmol*. 2003;121(11):1621-1624.
- Age-Related Eye Disease Study Research Group. A randomized, placebo-controlled, clinical trial of high-dose supplementation with vitamins C and E, beta carotene, and zinc for age-related macular degeneration and vision loss: AREDS report no. 8. *Arch Ophthalmol*. 2001;119(10):1417-1436.
- Bressler NM, Chang TS, Sufer IJ, et al; MARINA and ANCHOR Research Groups. Vision-related function after ranibizumab treatment by better—or worse-seeing eye: clinical trial results from MARINA and ANCHOR. *Ophthalmology*. 2010;117(4):747-56.e4.
- US Department of Commerce; United States Census Bureau. Statistical abstract of the United States: 2012. <https://www2.census.gov/library>
- /publications/2011/compendia/statab/131ed/2012-statab.pdf. Accessed August 18, 2017.
- Holz FG, Strauss EC, Schmitz-Valckenberg S, van Lookeren Campagne M. Geographic atrophy: clinical features and potential therapeutic approaches. *Ophthalmology*. 2014;121(5):1079-1091.
- Lim LS, Mitchell P, Seddon JM, Holz FG, Wong TY. Age-related macular degeneration. *Lancet*. 2012;379(9827):1728-1738.
- Lindblad AS, Lloyd PC, Clemons TE, et al; Age-Related Eye Disease Study Research Group. Change in area of geographic atrophy in the Age-Related Eye Disease Study: AREDS report number 26. *Arch Ophthalmol*. 2009;127(9):1168-1174.
- Tolentino MJ, Dennrick A, John E, Tolentino MS. Drugs in phase II clinical trials for the treatment of age-related macular degeneration. *Expert Opin Investig Drugs*. 2015;24(2):183-199.
- Burlina P, Freund DE, Dupas B, Bressler N. Automatic screening of age-related macular degeneration and retinal abnormalities. *Conf Proc IEEE Eng Med Biol Soc*. 2011;2011:3962-3966.
- Feeny AK, Tadarati M, Freund DE, Bressler NM, Burlina P. Automated segmentation of geographic atrophy of the retinal epithelium via random forests in AREDS color fundus images. *Comput Biol Med*. 2015;65:124-136.
- Freund DE, Bressler NM, Burlina P. Automated detection of drusen in the macula. Paper presented at: the Institute of Electrical and Electronics Engineers International Symposium on Biomedical Imaging; June 28-July 1, 2009; Boston, MA. <http://ieeexplore.ieee.org/document/5192983/>. Accessed August 18, 2017.
- Vapnik VN. *Statistical Learning Theory*. New York, NY: Wiley; 1998:416-417.
- Burlina P, Freund DE, Joshi N, Wolfson Y, Bressler NM. Detection of age-related macular degeneration via deep learning. Paper presented at: the Institute of Electrical and Electronics Engineers International Symposium on Biomedical Imaging; April 13-16, 2016; Prague, Czech Republic. <http://ieeexplore.ieee.org/document/7493240/>. Accessed August 18, 2017.
- Lowe DG. Distinctive image features from scale invariants keypoints. *Int J Comput Vis*. 2004;60(2):91-110. <https://doi.org/10.1023/B:VISI.0000029664.99615.94>
- Rajagopalan AN, Burlina P, Chellappa R. Detection of people in images. Paper presented at: the International Joint Conference on Neural Networks; July 10-16, 1999; Washington, DC. <http://www.ee.iitm.ac.in/~raju/conf/c15.pdf>. Accessed August 18, 2017.
- Trucco E, Ruggeri A, Karnowski T, et al. Validating retinal fundus image analysis algorithms: issues and a proposal. *Invest Ophthalmol Vis Sci*. 2013;54(5):3546-3559.
- Burlina P, Pacheco KD, Joshi N, Freund DE, Bressler NM. Comparing humans and deep learning performance for grading AMD: a study in using universal deep features and transfer learning for automated AMD analysis. *Comput Biol Med*. 2017;82:80-86.
- Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. <https://arxiv.org/abs/1311.2524>. Accessed August 18, 2017.
- Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. <https://papers.nips.cc/paper/4824-Imagenet-classification-with-deep-convolutional-neural-networks.pdf>. Accessed August 18, 2017.
- Razavian AS, Azizpour H, Sullivan J, Carlsson S. CNN features off-the-shelf: an astounding baseline for recognition. Paper presented at: the Institute of Electrical and Electronics Engineers Conference of Computer Vision and Pattern Recognition; May 12, 2014; Stockholm, Sweden. <https://arxiv.org/pdf/1403.6382.pdf>. Accessed August 18, 2017.
- Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. <https://arxiv.org/abs/1409.1556>. Accessed August 18, 2017.
- Szegedy C, Liu W, Yangqing J. Going deeper with convolutions. Paper presented at: the Institute of Electrical and Electronics Engineers Conference of Computer Vision and Pattern Recognition; June 7-12, 2015; Boston, MA. <http://ieeexplore.ieee.org/document/7298594/>. Accessed August 18, 2017.
- 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(22):2402-2410.
- Quellec G, Charrière K, Boudi Y, Cochener B, Lamard M. Deep image mining for diabetic retinopathy screening. *Med Image Anal*. 2017;39:178-193.
- Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. Paper presented at: the European Conference on Computer Vision; September 6-12, 2014; Zürich, Switzerland. <https://arxiv.org/abs/1311.2901>. Accessed August 18, 2017.
- Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y. Overfeat: integrated recognition, localization and detection using convolutional networks. <https://arxiv.org/abs/1312.6229>. Accessed August 18, 2017.
- Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174.
- Pacheco K, et al. Evaluation of automated drusen detection system for fundus photographs of patients with age-related macular degeneration. *Invest Ophthalmol Vis Sci*. 2016;57(12):1611. <http://iovs.arvojournals.org/article.aspx?articleid=2560240&resultClick=1>. Accessed August 18, 2017.