HHS Public Access

Author manuscript

Ophthalmology. Author manuscript; available in PMC 2023 May 01.

Published in final edited form as:

Ophthalmology. 2022 May ; 129(5): 571–584. doi:10.1016/j.ophtha.2021.12.017.

DeepLensNet: Deep Learning Automated Diagnosis and Quantitative Classification of Cataract Type and Severity

Tiarnan D.L. Keenan, BM BCh, PhD[1,*], Qingyu Chen, PhD[2,*], Elvira Agrón, MA[1], Yih-Chung Tham, PhD[3,4], Jocelyn Hui Lin Goh, PhD[3], Xiaofeng Lei, Msc[5], Yi Pin Ng, BSc[5], Yong Liu, PhD[4,5], Xinxing Xu, PhD[4,5], Ching-Yu Cheng, MD, PhD[3,4,5,6], Mukharram M. Bikbov, MD[7], Jost B. Jonas, MD[8,9,10], Sanjeeb Bhandari, MD, PhD[1], Geoffrey K. Broadhead, MD[1], Marcus H. Colyer, MD[11,12], Jonathan Corsini, MD[13], Chantal Cousineau-Krieger, MD[1], William Gensheimer, MD[14,15], David Grasic, MD[1], Tania Lamba, MD[16], M. Teresa Magone, MD[1], Michele Maiberger, MD[16], Arnold Oshinsky, MD[16], Boonkit Purt, MD[12,17], Soo Y. Shin, MD[16], Alisa T. Thavikulwat, MD[1], Zhiyong Lu, PhD[2], Emily Y. Chew, MD[1] AREDS Deep Learning Research Group

[1]Division of Epidemiology and Clinical Applications, National Eye Institute, National Institutes of Health, Bethesda, Maryland.

[2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland.

[3]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore.

Correspondence: Tiarnan D.L. Keenan, BM BCh, PhD, National Eye Institute, National Institutes of Health, 9000 Rockville Pike, Bethesda, MD 20892. tiarnan.keenan@nih.gov; Qingyu Chen, PhD, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894. qingyu.chen@nih.gov; Emily Y. Chew, MD, National Eye Institute, National Institutes of Health, 9000 Rockville Pike, Bethesda, MD 20892. echew@nei.nih.gov; Zhiyong Lu, PhD, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894. luzh@ncbi.nlm.nih.gov.

*T.D.L.K. and Q.C. contributed equally to this work.

Author Contributions:

Conception and design: Keenan, Chen, Agrón, Lu, Chew

Data collection: Keenan, Chen, Tham, Goh, Lei, Ng, Liu, Xu, Cheng, Bikbov, Jonas, Bhandari, Broadhead, Colyer, Corsini, Cousineau-Krieger, Gensheimer, Grasic, Lamba, Magone, Maiberger, Oshinsky, Purt, Shin, Thavikulwat, Lu, Chew

Analysis and interpretation: Keenan, Chen, Agrón, Tham, Goh, Lei, Ng, Liu, Xu, Cheng, Bikbov, Jonas, Lu, Chew

Obtained funding: N/A; Study was performed as part of the authors' regular employment duties. No additional funding was provided.

Overall responsibility: Keenan, Chen, Agrón, Tham, Goh, Lei, Ng, Liu, Xu, Cheng, Bikbov, Jonas, Bhandari, Broadhead, Colyer, Corsini, Cousineau-Krieger, Gensheimer, Grasic, Lamba, Magone, Maiberger, Oshinsky, Purt, Shin, Thavikulwat, Lu, Chew

Supplemental material available at www.aaojournal.org.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

This research was supported in part by the Intramural Research Program of the National Eye Institute, National Institutes of Health, Department of Health and Human Services (Bethesda, MD), and the National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health. The sponsor and funding organization participated in the design and conduct of the study; data collection, management, analysis, and interpretation; and the preparation, review and approval of the manuscript.

Dr. Chew, an associate editor of the *Journal*, was recused from the peer-review process of this article and had no access to information regarding its peer review.

The views expressed herein are those of the authors and do not reflect the official policy or position of Walter Reed National Military Medical Center, Madigan Army Medical Center, Joint Base Andrews, the U.S. Army Medical Department, the U.S. Army Office of the Surgeon General, the Department of the Air Force, the Department of the Army/Navy/Air Force, Department of Defense, the Uniformed Services University of the Health Sciences or any other agency of the U.S. Government. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

HUMAN SUBJECTS: Human subjects were included in this study. Institutional review board approval was obtained at each clinical site and written informed consent for the research was obtained from all study participants. All research adhered to the tenets of the Declaration of Helsinki.

No animal subjects were used in this study.

[4]Duke-NUS Medical School, Singapore.

[5]Institute of High Performance Computing, A*STAR, Singapore.

[6]Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore.

[7]Ufa Eye Research Institute, Ufa, Russia.

[8]Department of Ophthalmology, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany.

[9]Institute of Molecular and Clinical Ophthalmology, Basel, Switzerland.

[10]Privatpraxis Prof Jonas und Dr Panda-Jonas, Heidelberg, Germany.

[11]Department of Ophthalmology, Madigan Army Medical Center, Tacoma, Washington.

[12]Department of Surgery, Uniformed Services University of the Health Sciences, Bethesda, Maryland.

[13]Warfighter Eye Center, Malcolm Grow Medical Clinics and Surgery Center, Joint Base Andrews, Maryland.

[14]White River Junction Veterans Affairs Medical Center, White River Junction, Vermont.

[15]Geisel School of Medicine, Dartmouth, New Hampshire.

[16]Washington DC Veterans Affairs Medical Center, Washington, D.C.

[17]Department of Ophthalmology, Walter Reed National Military Medical Center, Bethesda, Maryland.

Abstract

Purpose: To develop deep learning models to perform automated diagnosis and quantitative classification of age-related cataract from anterior segment photographs.

Design: DeepLensNet was trained by applying deep learning models to the Age-Related Eye Disease Study (AREDS) dataset.

Participants: A total of 18 999 photographs (6333 triplets) from longitudinal follow-up of 1137 eyes (576 AREDS participants).

Methods: Deep learning models were trained to detect and quantify nuclear sclerosis (NS; scale 0.9–7.1) from 45-degree slit-lamp photographs and cortical lens opacity (CLO; scale 0%–100%) and posterior subcapsular cataract (PSC; scale 0%–100%) from retroillumination photographs. DeepLensNet performance was compared with that of 14 ophthalmologists and 24 medical students.

Main Outcome Measures: Mean squared error (MSE).

Results: On the full test set, mean MSE for DeepLensNet was 0.23 (standard deviation [SD], 0.01) for NS, 13.1 (SD, 1.6) for CLO, and 16.6 (SD, 2.4) for PSC. On a subset of the test set (substantially enriched for positive cases of CLO and PSC), for NS, mean MSE for DeepLensNet was 0.23 (SD, 0.02), compared with 0.98 (SD, 0.24; $P = 0.000001$) for the ophthalmologists and

1.24 (SD, 0.34; $P = 0.000005$) for the medical students. For CLO, mean MSE was 53.5 (SD, 14.8), compared with 134.9 (SD, 89.9; $P = 0.003$) for the ophthalmologists and 433.6 (SD, 962.1; $P = 0.0007$) for the medical students. For PSC, mean MSE was 171.9 (SD, 38.9), compared with 176.8 (SD, 98.0; $P = 0.67$) for the ophthalmologists and 398.2 (SD, 645.4; $P = 0.18$) for the medical students. In external validation on the Singapore Malay Eye Study (sampled to reflect the cataract severity distribution in AREDS), the MSE for DeepSeeNet was 1.27 for NS and 25.5 for PSC.

Conclusions: DeepLensNet performed automated and quantitative classification of cataract severity for all 3 types of age-related cataract. For the 2 most common types (NS and CLO), the accuracy was significantly superior to that of ophthalmologists; for the least common type (PSC), it was similar. DeepLensNet may have wide potential applications in both clinical and research domains. In the future, such approaches may increase the accessibility of cataract assessment globally. The code and models are available at https://github.com/ncbi/deeplensnet.

Keywords

Artificial intelligence; Automated diagnosis; Cataract; Cortical cataract; Deep learning; Nuclear sclerosis; Posterior subcapsular cataract; Severity classification; Telemedicine; Teleophthalmology

Cataract is the leading cause of legal blindness worldwide.[1,2] Its prevalence is predicted to increase further in the coming decades because of aging population demographics in many countries.[3-5] In its more advanced forms, cataract causes severe and typically bilateral visual impairment and requires surgical extraction and intraocular lens implantation.[1,6] Three main types of age-related cataract exist, related to the anatomic part of the crystalline lens affected by opacification: nuclear, cortical, and posterior subcapsular. Although the prevalence of all 3 types increases with age, each has partially distinct risk factors, visual symptoms, and rates of progression.[1,6,7]

The diagnosis and severity classification of cataract typically require in-person evaluation by an ophthalmologist.[1,6] This may limit the accessibility of cataract assessment, particularly for individuals living in remote areas or in countries with few ophthalmologists.[8-10] Even with cataract assessment by human experts in a research context, the accuracy and consistency of diagnosis and severity classification are known to be suboptimal[11-13]; this phenomenon is likely to be more marked in routine clinical practice. In the research domain, this makes it more difficult to quantify cataract progression accurately over time in interventional clinical trials and natural history studies and to compare consistently between studies. In the clinical domain, this inconsistency between (and even within) clinicians is typically accompanied by the use of less granular grading scales. Thus, detecting cataract progression and correctly attributing symptoms and acuity changes to cataract versus coexisting pathology may be less precise.

In this context, automated approaches to cataract diagnosis and classification have important potential advantages, including speed and accessibility, as well as accuracy and consistency.[11] They also lend themselves well to teleophthalmology approaches, where images can be obtained in a place and way that is safe and convenient for patients, with automated algorithms applied to these images.[14] In particular, deep learning has demonstrated high levels of performance in diagnosis and classification tasks in medicine,

including in ophthalmology, where diagnosis is often based on grading of anatomic features.[14-17]

The Age-Related Eye Disease Study (AREDS) was a multicenter, prospective study of the clinical course of age-related cataract and age-related macular degeneration (AMD).[18,19] The primary aim of the current study was to use the AREDS dataset to train and test deep learning models to perform diagnosis and severity classification of age-related cataract (including all 3 anatomic types) in a quantitative way that closely resembles human expert diagnosis and classification. An additional objective was to characterize human performance at 2 levels of experience to compare automated and human performance levels.

Methods

Dataset

The dataset used for deep learning model training and validation was the dataset of images, labels, and accompanying clinical information from the AREDS. The AREDS was a multicenter, prospective study to assess the clinical course, prognosis, and risk factors of age-related cataract and AMD, as well as a phase III randomized clinical trial designed to assess the effects of nutritional supplements on cataract and AMD progression.[18,19] For cataract, its primary outcome was the occurrence in at least 1 eye of progression in 1 or more of the 3 cataract types or cataract surgery.[19] Institutional review board approval was obtained at each clinical site, and written informed consent for the research was obtained from all study participants. The research was conducted under the tenets of the Declaration of Helsinki.

The AREDS study design has been described previously.[18,19] In short, at baseline and annual study visits, comprehensive eye examinations were performed by certified study personnel using a standardized protocol. The study visits included the capture of 3 types of anterior segment photograph for each eye, according to a standardized imaging protocol (Fig 1): (1) a slit-lamp photograph with the beam at an angle of 45° (width 0.3 mm and height 9.0 mm), bisecting the central lens, and focused near the center of the lens sulcus (Topcon SL-6E Photo Slit-Lamp Camera, Topcon Corporation); (2) a retroillumination photograph focused on the iris at the pupillary margin (Neitz Retroillumination Camera, Neitz Instruments Company, Ltd); and (3) a second retroillumination photograph focused on posterior subcapsular opacities (if present) or 3 to 5 mm posterior to the plane of the anterior photograph (if absent).

The dataset consisted of all AREDS anterior segment images where digitized images were available: 18999 images (consisting of 6333 image triplets, i.e., slit-lamp, anterior retroillumination, and posterior retroillumination photographs) from 1137 eyes of 576 participants. The dataset was split randomly into 3 sets, with the division made at the participant level (such that all images from a single participant were present in 1 of the 3 sets only): 70% for training, 10% for validation, and 20% for testing of the models. The characteristics of the participants and images used for training and testing are shown in Table 1 and Figure 2.

Ground Truth Labels

The ground truth labels used for model training and testing were the grades previously assigned to the images by expert human graders at the Wisconsin Reading Center (University of Wisconsin). The protocol and definitions used for cataract grading have been described previously.[20] In brief, the 45-degree slit-lamp photographs were used to grade nuclear cataract by comparison with 7 standard photographs of lenses with increasingly severe nuclear cataract (Fig 1A). A decimal grade ranging from 0.9 (less severe than Standard 1) to 7.1 (more severe than Standard 7) was assigned. The anterior retroillumination photographs were used to grade cortical cataract (Fig 1B), and the posterior retroillumination photographs were used to grade posterior subcapsular cataract (PSC) (Fig 1C) in terms of percentage area involvement. In both cases, any lens area that was definitely darkened (defined as a definite darkly shaded interruption of the reddish-orange fundus reflex) was considered involved, regardless of the density of the opacity. Percentage area involvement was calculated as follows: a grid, consisting of 2 concentric circles (with diameters 2 mm and 5 mm), was used to divide the photograph into 9 subfields. For both cortical cataract and PSC, considered separately, the main variable was the percentage area involvement of the 5-mm diameter circle occupied by definite opacity (calculated by combining the subfield percentages, weighted according to the size of each subfield). In addition, the percentage area involvement of the central 2-mm diameter circle was recorded as a separate variable because visual symptoms and acuity are expected to be affected particularly by cataract with central involvement. Intergrader and intragrader agreement rates were high, as described previously.[20]

Overall, 5 variables were considered in this study. The 3 main variables were nuclear cataract (nuclear sclerosis [NS]) (0.9–7.1), cortical lens opacity (CLO) (0.0%–100.0%), and PSC (0.0%–100.0%). The 2 secondary variables were central cortical cataract (CLO-center) (0.0%–100.0%) and central PSC (PSC-center) (0.0%–100.0%).

Deep Learning Framework

The proposed deep learning framework consists of 5 deep learning models, 1 for each of the 5 cataract classification variables. The NS model takes a 45-degree slit-lamp photograph as its input and predicts the NS variable. From an anterior retroillumination photograph as input, the CLO model predicts the CLO variable, and the CLO-center model predicts the CLO-center variable. Likewise, from a posterior retroillumination photograph, the PSC model predicts the PSC variable, and the PSC-center model predicts the PSC-center variable.

For all 5 tasks, a convolutional neural network (CNN)-based regression model was used. Given an input image, we first used a CNN (e.g., InceptionV3 or DenseNet) as a backbone to extract image features. The image features were further processed by an average pooling layer, fully connected layers, and a prediction layer to provide the quantitative output for severity classification. The CNN backbone was pretrained on ImageNet (an image database of >14 million natural images with corresponding labels, using methods described previously[21]) and then fine-tuned on the training set. We comparatively analyzed InceptionV3, ResNet, and DenseNet as the CNN backbone. These CNN models (e.g.,

number of layers and parameters) have been described in detail previously.[22] For the current work, InceptionV3 achieved the highest performance and was used as the backbone. The hyperparameter values used during the training process are summarized in Table 2. Each input image was resized from the original size (4008 × 2672 pixels) to 501 × 334 pixels. The model parameters were updated using the Adam optimizer (learning rate of 0.0001) for every minibatch of 16 images. An early stop procedure was applied to avoid overfitting: the training was stopped if the loss on the validation set no longer decreased for 5 epochs. In addition, image augmentation procedures were used, as follows, to increase the dataset size and strengthen model generalizability: (1) rotation (clockwise by 0° – 180°, selected randomly); (2) horizontal flip; and (3) vertical flip.

For each of the 5 tasks, we trained a deep learning model 10 times (with different random seeds each time) using the same training, validation, and test split shown in Table 1 to create 10 individual models (i.e., 50 models in total). The models were implemented using TensorFlow.[23] All experiments were conducted on a server with 48 Intel Xeon central processing units, using 3 NVIDIA Tesla V100 graphics processing units for training and testing, with 512 GB available in random access memory.

Evaluation of the Deep Learning Models

For each of the 5 variables, the models were evaluated against the gold standard reading center grades on the full test set of images. The primary performance metric calculated for each model was the mean squared error (MSE), given the quantitative nature of the task.[24] The MSE is calculated as the average of the squares of the errors, where the errors are the differences between the estimated values and the ground truth values from reading center grading. Because the errors are squared, the MSE is always positive, and much higher penalties are awarded for large errors (which corresponds well to a desired performance metric in a clinical setting). Thus, a lower MSE (close to zero) indicates better agreement with the ground truth values, and a higher MSE indicates worse agreement.

Evaluation of the Deep Learning Models in Comparison with Human Grading

For each of the 5 variables, the performance of the deep learning models was compared with the performance of 38 humans who manually graded the same images (when viewed on a computer screen at full image resolution), independently of each other. The 38 humans comprised 2 levels of specialization: 14 ophthalmologists (specifically 10 at attending level and 4 at fellowship level) and 24 medical students, where the ophthalmologists had experience in cataract grading in routine clinical practice and the medical students did not. Before grading, all of the human graders were provided with the same cataract grading definitions as those used by the reading center graders (i.e., as described previously).

For this evaluation, the test set of images was a subset of the full test set and comprised 100 45-degree slit-lamp images (for grading of NS), 100 anterior retroillumination images (for grading of CLO and CLO-center), and 100 posterior retroillumination images (for grading of PSC and PSC-center), all considered at the image level rather than the participant level (i.e., rather than requiring all 3 image types for any participant to be present). The 45-degree slit-lamp images were a random subset of the full test set. The anterior and

posterior retroillumination images were deliberately enriched for more severe cases, as the distributions of CLO severity and PSC severity in the full test set were both highly skewed toward negative cases. Specifically, the 100 anterior retroillumination images were selected as follows: 33 images with CLO 0.0%, 33 images with CLO 0.1% to 10.0%, and 34 images with CLO 10.1% to 100.0% (selected randomly in each of the 3 groups). The 100 posterior retroillumination images were selected using the same approach applied to PSC.

For each of the 5 variables, the human graders were evaluated against the gold standard reading center grades. Again, the primary performance metric was the MSE.[24] For each variable, the Mann–Whitney U (Wilcoxon rank-sum) test (2-tailed; 99% confidence intervals) was used to compare the performance of the (1) deep learning models, (2) ophthalmologists, and (3) medical students.

In post hoc analyses, the performance of both the models and the human graders was calculated according to pupil diameter (which had been measured by reading center graders). Pupil diameter was considered in the following categories: diameter <5 mm, 5 to 7 mm, and >7 mm. The threshold of 5 mm was chosen because the reading center gradings of CLO and PSC were performed on the 5-mm diameter circle, as described earlier.

Attention Maps

Attention maps were generated to investigate the image locations that contributed most to the decision making by the deep learning models. This was done by back-projecting the last convolutional layer of the neural network. The keras-vis package was used to generate the attention maps.[25] This was done for a sample of 40 images for each of the 3 cataract types (i.e., 120 images in total) from the full test set. The samples were selected as follows to obtain a representative sample of both positive and negative cases for each cataract type: for cortical cataract, a random sample of 20 positive images (CLO >0%, according to reading center grading) and 20 negative images (CLO = 0%); for PSC, a random sample of 20 positive images (PSC >0%) and 20 negative images (PSC = 0%); for nuclear cataract, a random sample of 20 images (any severity) and 20 images with more severe cataract (selected randomly from the 100 images with the highest NS gradings).

External Validation

The dataset used for external validation of the deep learning models was the Singapore Malay Eye Study (SiMES). This study has been described previously.[26] In brief, it was a cross-sectional population-based study of eye disease in adult individuals (aged 40–80 years) of Malay ethnicity in Singapore. Participants underwent standardized assessment that included digital lens imaging; 45-degree slit-lamp images were obtained using the Topcon DC-1 camera, and black-and-white retroillumination images were obtained using the Nidek EAS-1000 camera. These images were graded for NS (following the Wisconsin Cataract Grading System, range 1–5[27]) and for cortical cataract and PSC (each following the Wisconsin Cataract Grading System, range 0%–100%). The prevalence of age-related cataract in this dataset, overall and by subtype, has been reported previously.[28]

For NS, of the 4669 45-degree slit-lamp images with grades available, a subset of 200 images was selected randomly to correspond to the NS severity distribution in the

AREDS dataset for meaningful comparison of the results. Likewise, for PSC, of the 2266 retroillumination images with grades available, a subset of 200 images was selected randomly to correspond to the PSC severity distribution in the AREDS dataset for the same reason. For cortical cataract, the image type was not compatible with that used from AREDS to train the models.

The external validation performance of the models was evaluated against the gold standard grades from the SiMES dataset. For NS, the SiMES gold standard grades were first converted from the Wisconsin 1 to 5 scale to the AREDS 1 to 7 scale based on 3 of the standard photographs being shared between the 2 scales.[20,27] No conversion was required for the posterior subcapsular grades because the same 0% to 100% grading scale was used for both. The same performance metric was used as that for internal testing, that is, MSE.

Results

Automated Classification of Cataract Severity by Deep Learning

For the automated classification of cataract severity by the deep learning models, the results on the full test set are shown in Table 3. For NS, with NS grading considered on the 0.9 to 7.1 scale, the mean MSE for the 10 deep learning models was 0.23 (standard deviation [SD], 0.01). For CLO, with CLO grading considered as percentages, the mean MSE was 13.1 (SD, 1.6). For PSC, with PSC grading considered as percentages, the mean MSE was 16.6 (SD, 2.4). The small SDs suggested a high level of consistency between the 10 models trained for each task. For cortical and PSC, the 2 secondary variables (COL-center and PSC-center) had a mean MSE of 53.7 (SD, 4.9) and 51.9 (SD, 6.5), respectively.

Performance of Deep Learning Models in Comparison with Human Graders

For the classification of cataract severity by the deep learning models versus the human graders, the results on the subset of the test set are shown in Figure 3 and Table 4. For NS and CLO, the performance of the deep learning models was significantly superior to that of the 14 ophthalmologists. Likewise, it was significantly superior to that of the 24 medical students. For NS, the mean MSE for the 10 deep learning models was 0.23 (SD, 0.02), compared with 0.98 (SD, 0.24; $P = 0.000001$) for the ophthalmologists and 1.24 (SD, 0.34; $P = 0.000005$) for the medical students. For CLO, the mean MSE for the models was 53.5 (SD, 14.8), compared with 134.9 (SD, 89.9; $P = 0.003$) for the ophthalmologists and 433.6 (SD, 962.1; $P = 0.0007$) for the medical students. For PSC, the performance of the deep learning models was similar to that of the ophthalmologists and numerically but not significantly superior to that of the medical students. Specifically, the mean MSE for the models was 171.9 (SD, 38.9), compared with 176.8 (SD, 98.0; $P = 0.67$) for the ophthalmologists and 398.2 (SD, 645.4; $P = 0.18$) for the medical students. For CLO-center and PSC-center, the performance of the deep learning models was numerically slightly inferior to that of the ophthalmologists and numerically superior to that of the medical students; however, it was not significantly different from either ($P = 0.31$ and 0.56, respectively, for CLO-center, and $P = 0.23$ and 0.87, respectively, for PSC-center).

Performance of Deep Learning Models According to Pupil Diameter

The results of the performance of the deep learning models on the full test set, according to pupil diameter, are shown in Table S1 (available at www.aaojournal.org), and the results of the performance of the deep learning models and the human graders on the subset of the test set, according to pupil diameter, are shown in Table S2 (available at www.aaojournal.org). On the full test set, for all 5 cataract variables, the performance of the deep learning models was relatively similar for eyes with smaller (5–7 mm) versus larger (>7 mm) pupil diameter. For NS, mean MSE was extremely similar; for CLO, it was numerically slightly worse in cases with smaller pupil diameter; for PSC, it was numerically better. On the subset of the test set, for NS, mean MSE was again extremely similar; for CLO, it was numerically better in cases with smaller pupil diameter, and for PSC, it was numerically worse, although the numbers with smaller pupils were low. For the 14 ophthalmologists, for NS and CLO, mean MSE was numerically worse in cases with smaller pupil diameter (particularly for NS), and for PSC, it was similar.

Attention Maps

Attention maps were generated and superimposed on the images. For each image, these demonstrate quantitatively the relative contributions made by each pixel to the grading prediction. Representative examples of these attention maps are shown for each of the 3 cataract types in Figure 4. The full set of 120 attention maps is available at https://github.com/ncbi/deeplensnet. For positive cases of CLO or PSC, the areas of high signal (that contributed most to the grading prediction) seemed to correspond closely to the location of the relevant opacity, as observable to human graders. This is despite the fact that the algorithms were not subject to any supervision or spatial guidance. By contrast, for the negative cases, no areas of high signal were observed in the distribution of the lens. In addition, the shape and extent of the areas of high signal seemed to correspond well with those of the opacities; for example, they differed in images with a single plaque of opacity versus widespread opacity in a particular distribution (Fig 4). This is consistent with the ability of the algorithms to perform quantitative grading. Similar findings were observed for nuclear cataract. In general, for moderate and severe cases, a single area of high signal was located at the lens nucleus; for absent or mild cases, no areas of high signal were observed there. Thus, through these attention maps, the outputs of the deep learning models seem to display a degree of face validity and interpretability for the detection and classification of cataract.

External Validation on the SiMES Dataset

For NS, with grading considered on the 0.9 to 7.1 scale (after conversion of the gold standard grades from the Wisconsin scale to the AREDS scale), the MSE for the NS deep learning model was 1.27. For PSC, with grading considered as percentages, the MSE for the PSC deep learning model was 25.5.

Discussion

Main Findings and Interpretation

The deep learning framework achieved automated and quantitative classification of cataract severity with a high level of accuracy for all 3 types of age-related cataract. On the full test set, the MSE was very low for each of the 3 types. A subset of the full test set was designed as a very challenging test set by deliberately enriching for positive cases of cortical and PSC. On this, performance remained high for the 2 most common types of age-related cataract (nuclear and cortical) and remained moderately high for the other type (posterior subcapsular). The latter result likely relates to the lower number of positive cases of PSC in the training set, which reflects the lower prevalence in the population. The performance of the deep learning framework was significantly superior to that of ophthalmologists for the 2 most common types of cataract (nuclear and cortical). For the least common type (posterior subcapsular), the accuracy was similar to that of ophthalmologists and numerically superior to that of medical students. The performance was not markedly lower in eyes with smaller pupils. For NS grading, accurate grading seemed possible irrespective of pupil size. This differed from the situation with human grading of NS, as ophthalmologist performance seemed to be worse in eyes with smaller pupils.

The subset of the full test set demonstrated validity in its discriminative power in that a substantial difference in classification performance was observed between human graders with and without ophthalmology experience. Interestingly, this difference was particularly apparent for cortical and PSC and less marked for nuclear cataract. This may relate to the former 2 being more difficult tasks for inexperienced graders. In the former 2 cases, both based on retroillumination photographs, the grader must decide on the presence of opacities in the image (passing the threshold of definitely darkened), distinguish between cortical and posterior subcapsular opacity, and add up the affected areas to derive the final percentage. By comparison, the grading of nuclear cataract may be easier for inexperienced graders; the task is simpler in that grading involves only 1 step and is performed by direct comparison with 7 standard photographs.

For cortical and PSC, the 2 secondary variables (COL-center and PSC-center) had lower performance metrics. This is likely related to both the higher difficulty level of these tasks and the lower number of positive instances in the training dataset. Deep learning performance was numerically superior to that of the medical students but numerically slightly inferior to that of the ophthalmologists.

Clinical Importance and Implications

The ability of a deep learning framework to perform automated cataract diagnosis on an ordinal scale, with high accuracy, consistency, and throughput, means that it may have broad applications, particularly in the research domain. The highly quantitative grading (separately by cataract subtype) may be valuable for applications where granular and accurate grading is required, for example, interventional clinical trials, natural history studies, and epidemiological studies. Likewise, the objectivity and consistency may be helpful in scenarios where these are important, for example, for consistency between

different studies; however, this quantitative grading also can be easily converted into one of multiple other grading systems, depending on the particular application, including less granular scales (e.g., NS 0–4+) often used in routine clinical practice.

In the research domain, potential applications may include (1) cross-sectional epidemiological studies of cataract prevalence and risk factors, (2) longitudinal natural history studies and risk factors for progression, and (3) interventional clinical trials. For the former, a previous study showed that even with the technology available in 1999, a digital cataract grading system seemed to be more cost-effective than human grading for a large epidemiological study, as well as more accurate and consistent.[29] For the latter, the high accuracy and consistency of the quantitative grading mean that clinical trials could remain highly powered despite smaller size and shorter duration.

In the clinical domain, potential applications in the future could include (1) cataract screening in primary care; (2) cataract screening alongside diabetic retinopathy screening; (3) more objective and quantitative diagnosis and grading in secondary/tertiary care; and (4) utility in surgical decision making, planning, case allocation, and risk stratification/prediction.

For the latter, if validated, such algorithms could assist ophthalmologists and patients in decisions around the risk/benefit balance of pursuing surgery or in surgical planning (e.g., more accurate assessments of what phacoemulsification power may be required for a particular case). A previous study of the Lens Opacities Classification System III demonstrated an exponential increase in phacoemulsification energy as nuclear cataract grades increased.[30] As argued previously, grading systems like this can be useful tools in creating operative plans, improving the allocation of appropriate cases to surgeons with the appropriate experience, and providing more accurate predictions of surgical risks and visual outcomes for patients and physicians.[11] Implementing cataract surgery risk stratification systems is thought to decrease complication rates, and nuclear cataract density is a key factor in all risk stratification systems.[31,32] Therefore, improving the accuracy and consistency of density assessments should lead to improved accuracy of risk stratification.

In the future, deep learning models are likely to be able to detect other important risk factors, such as pseudoexfoliation or zonular dehiscence, small pupils, and corneal endothelial disease.[33] Together with quantitative cataract grading, these outputs could be combined to generate more accurate surgical risk calculators (e.g., for posterior capsule rupture,[34] corneal decompensation,[35,36] and cystoid macular edema) that assist with tasks including risk/benefit decisions, case allocation, and operative planning. Overall, more accurate, consistent, and granular cataract grading should improve the performance of these tasks and make them more evidence-based and comparable between centers, even if they are currently performed by many experienced surgeons in an ad hoc way.

Deep learning frameworks like this may also have potential applications in global ophthalmology, although several steps may be required for implementation in this setting. Cataract remains the leading cause of blindness worldwide, particularly in low- and middle-income countries, despite the fact that it is a reversible cause of vision loss and cataract

surgery is considered highly cost-effective.[1,2,30,31] Two main barriers to restoring vision exist in these cases: awareness of diagnosis and access to surgery.[37,38] Traditionally, both have required in-person access to an ophthalmologist, which can be problematic for individuals living in areas that are remote or have few ophthalmologists or for where evaluation is expensive.

The development of automated approaches that can diagnose cataract from images carries advantages. In the wider context of telemedicine and teleophthalmology, these potential benefits have been described in detail.[39,40] Images may be obtained in a place and way that is safe, convenient, and less expensive for patients, without the need for an ophthalmologist. For example, recent studies have demonstrated the utility of smartphone-based portable slit-lamp devices.[41-43] In this way, patients with cataract may be diagnosed accurately and rapidly, with potentially decreased travel and expenditure. Thus, approaches like this could substantially increase the accessibility of cataract diagnosis without compromising on accuracy. Clearly, symptomatic patients diagnosed with cataract would still require access to ophthalmologists for surgical treatment; however, a recent study in a low-/middle-income country has shown that teleophthalmology approaches can substantially increase the subsequent attendance of patients with confirmed eye conditions at ophthalmic hospitals.[44]

A technician would still be required to administer the dilating drops and take the photographs. A camera whose image output had been validated against the deep learning framework would also be needed. Thus, in the future, external validation using a low-cost camera in a global setting will be important. This is another important reason why the trained models and code are being made freely available, that is, to expedite and spread these efforts as widely as possible. It also means that other groups can use these models and code as a starting place to train and test their own deep learning models for a particular camera or image type or for a specific application or setting, even with smaller image datasets (by fine-tuning training). Automated cameras are now available that can take illuminated anterior segment photographs, that is, with automated alignment, focus, and image acquisition.[45] The combination of an automated camera and automated cataract grading would work particularly well in decreasing dependence on trained technicians or photographers; however, the cost of such devices means that they are likely not appropriate for use in low-income countries at present. Finally, in low-income countries, detection of severe cataract is already possible using visual acuity measurement and a penlight; however, approaches like this are likely to be poorly accurate in moderate cases or in cases where visual acuity may be decreased partly through separate pathology.

External Validation

External validation was possible for 2 of the main types of age-related cataract, NS and PSC, despite the almost complete absence of publicly available datasets, through international collaboration. Finding a dataset of cortical cataract images compatible with the image type used in the current study has not been possible yet but remains an area of active research. The dataset used for external validation was from a different ethnic origin (specifically adults of Malay ethnicity in Singapore), which represents a higher bar to pass than external validation on an ethnically similar population. Other important considerations include the

following: for both NS and PSC, the camera models used differed between the AREDS training images and the SiMES testing images; for PSC, the SiMES images were black and white, and the AREDS training images were color; no image preprocessing was conducted to account for these differences; no algorithm fine-tuning or retraining, even on a small sample of images, was performed before external validation; the gold standard SiMES grades were from a single human grader, whereas the AREDS gold standard grades were from reading center grading, and interpretation of the images and grading scales may differ partially between the 2; in particular, SiMES NS severity was graded following the Wisconsin 1 to 5 scale, whereas AREDS NS severity was graded on the AREDS 1 to 7 scale, so conversion of the gold standard labels was required; although 3 of the standard photographs are shared between the 2 scales,[20,27] the correspondence between the scales at other points is more uncertain, which makes exact conversion impossible and tends to penalize the deep learning models.

Despite these difficulties, on an external dataset designed to reflect the cataract severity distribution in AREDS (for more meaningful comparison of results), performance was respectable. In future research, we plan to investigate the potential for further improvements on the basis of fine-tuning training (i.e., on a small number of images of the external dataset of interest). Given our commitment to make the code and models publicly available, the same method could be used by other groups on any existing and future datasets obtained by different cameras in different settings. Using methods such as federated learning, training models on multiple datasets from different institutions without sharing images will be possible in the future; a global deep learning model with high performance and generalizability can be created by combining many locally trained models.[46]

Comparison with Literature

Relatively few previous studies have used artificial intelligence methods to perform automated grading of cataract from anterior segment images. These studies have been examined in several review articles.[47-49] Indeed, previous authors have noted recently that, compared with other common major age-related eye diseases, “AI [artificial intelligence] development in the domain of cataract is still relatively underexplored.”[48] This may be particularly surprising because cataract seems to lend itself very well to automated approaches. Unlike with some other age-related eye diseases, the spectrum of phenotypic expression in cataract is relatively narrow. Cataract appears in a similar way at the same anatomic locations, irrespective of ethnicity or other demographic differences, such that issues around generalizability to other populations may be easier to overcome. In addition, cataract can be diagnosed and classified purely on the anatomic appearance, without additional information (e.g., age or history of diabetes mellitus). Finally, it can be diagnosed on simple imaging modalities (e.g., not requiring complex 3-dimensional modalities like OCT, as with many retinal diseases). Thus, the small number of previous studies may relate more to the relative paucity of large datasets of anterior segment images with high-quality grading. In a recent global review of all publicly available datasets of ophthalmic imaging, the proportion related to cataract was only 4%; only 1 dataset contained slit-lamp photographs of eyes with cataract, and this contained just 60 images.[50]

Of the previous reports of artificial intelligence approaches to cataract diagnosis that do exist,[51-55] the majority[51-53] used traditional machine learning rather than the deep learning approaches that have been associated with improved performance. As expected, therefore, the performance in these reports was relatively modest. In addition, the majority performed grading of nuclear cataract only.[52-54] Indeed, previous authors have argued that most artificial intelligence approaches have "focused on a single specific cataract subtype, which can severely limit [their] application in real-world health-care settings."[49] For example, Cheung et al[52] reported an intraclass correlation coefficient of 0.81 for agreement between their automated approach and manual grading on the Wisconsin grading scale (0.1–5.0) for grading of nuclear cataract; however, the approach was only semiautomated because the segmentation failed and required manual correction in 5% of cases. Xu et al[53] reported agreement of 0.85 (for agreement within 0.5 of a grade) in the grading of nuclear cataract.

To our knowledge, the largest previous study was conducted in China.[55] It was based on 38 000 anterior segment images, although these were a mixture of transverse slit beam and diffuse illumination images (i.e., not by retroillumination) and of mydriatic and nonmydriatic images. Of note, the algorithm did not attempt to perform severity classification separately for the 3 types of cataract. In a staged approach, the first algorithm performed binary classification of cataract presence or absence, with an area under the receiver operating characteristic curve of 0.999. For images with cataract, the second algorithm performed binary classification of nuclear cataract only (based on grades I–II vs. III–IV on the Lens Opacities Classification System II), with an area under the receiver operating characteristic curve of 0.992 for mydriatic transverse slit-beam images. For images with grades I and II nuclear cataract, the third algorithm performed binary classification of other cataracts (presumably cortical or posterior subcapsular), with an area under the receiver operating characteristic curve of 0.949 on unspecified image types. Thus, the approach was still based principally on nuclear cataract and was essentially binary rather than quantitative in nature. Given these differences, no direct comparison can easily be made with the results in the current study. Finally, none of these previous studies compared performance of the automated approach with that of ophthalmologists.

Strengths, Limitations, and Future Work

The strengths of this study include analysis of all 3 types of age-related cataract. Regarding the ground truth labels, the study benefitted from centralized grading of all images by expert graders at a single reading center with standardized grading definitions; previous reports have validated this reading center grading, with high rates of intergrader agreement.[20] The quantitative nature of the grading achieved has advantages over binary assessments. The dataset for training and testing comprised a wide breadth of data because the images were drawn from many different clinics across the United States. This increased the likelihood of generalizability; in external validation on a dataset from a different ethnicity and study setting, this was reflected in respectable performance for grading of NS and PSC. Additional strengths include comparison with human performance (which has not been assessed in previous studies), using a large number of ophthalmologists.

The limitations include the number of images available for training because only a subset of all AREDS anterior segment photographs have been digitized. This is more relevant for cortical and PSC because of the lower number of positive cases. Although the performance of the PSC model was not as high as the other 2, it was still similar to that of ophthalmologists. The level of performance observed in this study may relate partly to the use of trained photographers with specific camera types in a clinical trial setting. This is partially addressed by the external validation, which was conducted using images from different camera types and a different population. In future work, we aim to move from external validation of the models to additional training on datasets from multiple institutions, ethnicities, and camera types; ideally, this could even include those obtained by smartphone-based portable slit-lamps, which can take transverse beam and retroillumination photographs[41-43]; however, performance might be lower with images acquired by technicians with a lower level of training in a routine clinical setting or a global ophthalmology setting. Ideally, each setting and use case would require separate testing in prospective studies.

In conclusion, we developed a deep learning framework, DeepLensNet, for the detailed assessment of age-related cataract. DeepLensNet was able to perform automated, accurate, and quantitative classification of cataract severity for all 3 types of age-related cataract. For nuclear and cortical cataract, the 2 most common types, the accuracy was significantly superior to that of ophthalmologists; for PSC, the least common type, the accuracy was similar. External validation on a dataset from a population of different ethnicity demonstrated acceptable performance for NS and PSC, despite differences in study setting and camera models. We are making the code and pretrained models available, for research use only, at https://github.com/ncbi/deeplensnet. In this way, we aim to maximize the transparency and reproducibility of this study and to provide a benchmark method for the further refinement and development of methodologies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Appendix

AREDS Deep Learning Research Group:

Priscilla Ajilore, Alex Akman, Nadim S. Azar, William S. Azar, Bryan Chan, Victor Cox, Amisha D. Dave, Rachna Dhanjal, Mary Donovan, Maureen Farrell, Francisca Finkel, Timothy Goblirsch, Wesley Ha, Christine Hill, Aman Kumar, Kristen Kent, Arielle Lee, Pujan Patel, David Peprah, Emma Piliponis, Evan Selzer, Benjamin Swaby, Stephen Tenney, and Alexander Zeleny.

Abbreviations and Acronyms:

AMD	age-related macular degeneration
AREDS	Age-Related Eye Disease Study

CLO	cortical lens opacity
CNN	convolutional neural network
MSE	mean squared error
NS	nuclear sclerosis
PSC	posterior subcapsular cataract
SD	standard deviation
SiMES	Singapore Malay Eye Study

References

1. Lam D, Rao SK, Ratra V, et al. Cataract. Nat Rev Dis Primers. 2015;1:15014. [PubMed: 27188414]
2. GBD 2019 Blindness and Vision Impairment Collaborators; Vision Loss Expert Group of the Global Burden of Disease Study. Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study. Lancet Glob Health. 2021;9:e144–e160. [PubMed: 33275949]
3. Hashemi H, Pakzad R, Yekta A, et al. Global and regional prevalence of age-related cataract: a comprehensive systematic review and meta-analysis. Eye (Lond). 2020;34:1357–1370. [PubMed: 32055021]
4. Flaxman SR, Bourne RRA, Resnikoff S, et al. Global causes of blindness and distance vision impairment 1990-2020: a systematic review and meta-analysis. Lancet Glob Health. 2017;5:e1221–e1234. [PubMed: 29032195]
5. Bastawrous A, Suni AV. Thirty year projected magnitude (to 2050) of near and distance vision impairment and the economic impact if existing solutions are implemented globally. Ophthalmic Epidemiol. 2020;27:115–120. [PubMed: 31810404]
6. Liu YC, Wilkins M, Kim T, et al. Cataracts. Lancet. 2017;390:600–612. [PubMed: 28242111]
7. Chang JR, Koo E, Agron E, et al. Risk factors associated with incident cataracts and cataract surgery in the Age-related Eye Disease Study (AREDS): AREDS report number 32. Ophthalmology. 2011;118:2113–2119. [PubMed: 21684602]
8. Resnikoff S, Lansingh VC, Washburn L, et al. Estimated number of ophthalmologists worldwide (International Council of Ophthalmology update): will we meet the needs? Br J Ophthalmol. 2020;104:588–592. [PubMed: 31266774]
9. Burton MJ, Ramke J, Marques AP, et al. The Lancet Global Health Commission on Global Eye Health: vision beyond 2020. Lancet Glob Health. 2021;9:e489–e551. [PubMed: 33607016]
10. Bastawrous A, Dean WH, Sherwin JC. Blindness and visual impairment due to age-related cataract in sub-Saharan Africa: a systematic review of recent population-based studies. Br J Ophthalmol. 2013;97:1237–1243. [PubMed: 23696652]
11. Gali HE, Sella R, Afshari NA. Cataract grading systems: a review of past and present. Curr Opin Ophthalmol. 2019;30:13–18. [PubMed: 30489359]
12. Chew EY, Kim J, Sperduto RD, et al. Evaluation of the Age-Related Eye Disease Study clinical lens grading system AREDS report No. 31. Ophthalmology. 2010;117:2112–2119 e2113. [PubMed: 20561686]
13. Tan AC, Wang JJ, Lamoureux EL, et al. Cataract prevalence varies substantially with assessment systems: comparison of clinical and photographic grading in a population-based study. Ophthalmic Epidemiol. 2011;18:164–170. [PubMed: 21780875]
14. Li JO, Liu H, Ting DSJ, et al. Digital technology, telemedicine and artificial intelligence in ophthalmology: a global perspective. Prog Retin Eye Res. 2020:100900. [PubMed: 32898686]

15. Esteva A, Chou K, Yeung S, et al. Deep learning-enabled medical computer vision. NPJ Digit Med. 2021;4:5. [PubMed: 33420381]
16. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. 2019;1:e271–e297. [PubMed: 33323251]
17. Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol. 2019;103:167–175. [PubMed: 30361278]
18. Age-Related Eye Disease Study Research Group. The Age-Related Eye Disease Study (AREDS): design implications. AREDS report no. 1. Control Clin Trials. 1999;20:573–600. [PubMed: 10588299]
19. Age-Related Eye Disease Study Research Group. A randomized, placebo-controlled, clinical trial of high-dose supplementation with vitamins C and E and beta carotene for age-related cataract and vision loss: AREDS report no. 9. Arch Ophthalmol. 2001;119:1439–1452. [PubMed: 11594943]
20. Age-Related Eye Disease Study Research Group. The Age-Related Eye Disease Study (AREDS) system for classifying cataracts from photographs: AREDS report no. 4. Am J Ophthalmol. 2001;131:167–175. [PubMed: 11228291]
21. Peng Y, Dharssi S, Chen Q, et al. DeepSeeNet: a deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs. Ophthalmology. 2019;126:565–575. [PubMed: 30471319]
22. Keenan TDL, Chen Q, Peng Y, et al. Deep learning automated detection of reticular pseudodrusen from fundus auto-fluorescence images or color fundus photographs in AREDS2. Ophthalmology. 2020;127:1674–1687. [PubMed: 32447042]
23. Abadi M, Agarwal A, Brevdo E, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. https://arxiv.org/abs/1603.04467; 2015. Accessed June 15, 2021.
24. Lin LI, Heyadat AS, Sinha B, Yang M. Statistical methods in assessing agreement: models, issues and tools. J Am Stat Assoc. 2002;97:257–270.
25. Kotikalapudi R. Kerasvis. https://github.com/raghakot/kerasvis; 2017. Accessed June 15, 2021.
26. Majithia S, Tham YC, Chee ML, et al. Cohort profile: The Singapore Epidemiology of Eye Diseases study (SEED). Int J Epidemiol. 2021;50:41–52. [PubMed: 33393587]
27. Klein BE, Klein R, Linton KL, et al. Assessment of cataracts from photographs in the Beaver Dam Eye Study. Ophthalmology. 1990;97:1428–1433. [PubMed: 2255515]
28. Tan AG, Tay WT, Mitchell P, et al. Prevalence of lens opacities in Asian Malays. Ophthalmic Epidemiol. 2012;19:380–387. [PubMed: 23171207]
29. Dimock J, Robman LD, McCarty CA, Taylor HR. Cost-effectiveness of digital cataract assessment. Aust N Z J Ophthalmol. 1999;27:208–210. [PubMed: 10484193]
30. Davison JA, Chylack LT. Clinical application of the lens opacities classification system III in the performance of phacoemulsification. J Cataract Refract Surg. 2003;29:138–145. [PubMed: 12551681]
31. Muhtaseb M, Kalhoro A, Ionides A. A system for preoperative stratification of cataract patients according to risk of intraoperative complications: a prospective analysis of 1441 cases. Br J Ophthalmol. 2004;88:1242–1246. [PubMed: 15377542]
32. Pooprasert P, Hansell J, Young-Zvandasara T, Muhtaseb M. Can applying a risk stratification system, preoperatively, reduce intraoperative complications during phacoemulsification? Curr Eye Res. 2021;46:318–323. [PubMed: 32730130]
33. Eleiwa T, Elsawy A, Ozcan E, Abou Shousha M. Automated diagnosis and staging of Fuchs' endothelial cell corneal dystrophy using deep learning. Eye Vis (Lond). 2020;7:44. [PubMed: 32884962]
34. Sparrow JM, Taylor H, Qureshi K, et al. The cataract national data set electronic multi-centre audit of 55,567 operations: case-mix adjusted surgeon's outcomes for posterior capsule rupture. Eye (Lond). 2011;25:1010–1015. [PubMed: 21546922]
35. Sorrentino FS, Bonifazzi C, Parmeggiani F, Perri P. A pilot study to propose a "harm scale", a new method to predict risk of harm to the corneal endothelium caused by longitudinal

phacoemulsification, and the subsequent effect of endothelial damage on post operative visual acuity. PLoS One. 2016;11:e0146580. [PubMed: 26761198]

36. Doors M, Berendschot TT, Touwslager W, et al. Phacopower modulation and the risk for postoperative corneal decompensation: a randomized clinical trial. JAMA Ophthalmol. 2013;131:1443–1450. [PubMed: 24030086]
37. Mailu EW, Virendrakumar B, Bechange S, et al. Factors associated with the uptake of cataract surgery and interventions to improve uptake in low- and middle-income countries: a systematic review. PLoS One. 2020;15:e0235699. [PubMed: 32645065]
38. Zhang XJ, Jhanji V, Leung CK, et al. Barriers for poor cataract surgery uptake among patients with operable cataract in a program of outreach screening and low-cost surgery in rural China. Ophthalmic Epidemiol. 2014;21:153–160. [PubMed: 24754232]
39. Parikh D, Armstrong G, Liou V, Husain D. Advances in telemedicine in ophthalmology. Semin Ophthalmol. 2020;35:210–215. [PubMed: 32644878]
40. Nuzzi R, Bovone D, Maradei F, et al. Teleophthalmology service: organization, management, actual current applications, and future prospects. Int J Telemed Appl. 2021;2021:8876957. [PubMed: 34188678]
41. Hu S, Wu H, Luan X, et al. Portable handheld slit-lamp based on a smartphone camera for cataract screening. J Ophthalmol. 2020;2020:1037689. [PubMed: 32832134]
42. Dutt S, Vadivel SS, Nagarajan S, et al. A novel approach to anterior segment imaging with smartphones in the COVID-19 era. Indian J Ophthalmol. 2021;69:1257–1262. [PubMed: 33913872]
43. Yazu H, Shimizu E, Okuyama S, et al. Evaluation of nuclear cataract with smartphone-attachable slit-lamp device. Diagnostics (Basel). 2020;10:576.
44. Rono H, Bastawrous A, Macleod D, et al. Effectiveness of an mHealth system on access to eye health services in Kenya: a cluster-randomised controlled trial. Lancet Digit Health. 2021;3(7):e414–e424. 10.1016/S2589-7500(21)00083-2. [PubMed: 34167763]
45. Topcon. Maestro2. https://topconhealthcare.com/products/maestro2/. Accessed November 23, 2021.
46. Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: concept and applications. ACM Trans Intell Syst Technol. 2019;10. Article 12.
47. Ting DSJ, Foo VH, Yang LWY, et al. Artificial intelligence for anterior segment diseases: emerging applications in ophthalmology. Br J Ophthalmol. 2021;105:158–168. [PubMed: 32532762]
48. Goh JHL, Lim ZW, Fang X, et al. Artificial intelligence for cataract detection and management. Asia Pac J Ophthalmol (Phila). 2020;9:88–95. [PubMed: 32349116]
49. Wu X, Liu L, Zhao L, et al. Application of artificial intelligence in anterior segment ophthalmic diseases: diversity and standardization. Ann Transl Med. 2020;8:714. [PubMed: 32617334]
50. Khan SM, Liu X, Nath S, et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. Lancet Digit Health. 2021;3:e51–e66. [PubMed: 33735069]
51. Acharya RU, Yu W, Zhu K, et al. Identification of cataract and post-cataract surgery optical images using artificial intelligence techniques. J Med Syst. 2010;34:619–628. [PubMed: 20703916]
52. Cheung CY, Li H, Lamoureux EL, et al. Validity of a new computer-aided diagnosis imaging program to quantify nuclear cataract from slit-lamp photographs. Invest Ophthalmol Vis Sci. 2011;52:1314–1319. [PubMed: 21051727]
53. Xu Y, Gao X, Lin S, et al. Automatic grading of nuclear cataracts from slit-lamp lens images using group sparsity regression. Med Image Comput Comput Assist Interv. 2013;16(Pt 2):468–475. [PubMed: 24579174]
54. Gao X, Lin S, Wong TY. Automatic feature learning to grade nuclear cataracts based on deep learning. IEEE Trans Biomed Eng. 2015;62:2693–2701. [PubMed: 26080373]
55. Wu X, Huang Y, Liu Z, et al. Universal artificial intelligence platform for collaborative management of cataracts. Br J Ophthalmol. 2019;103:1553–1560. [PubMed: 31481392]

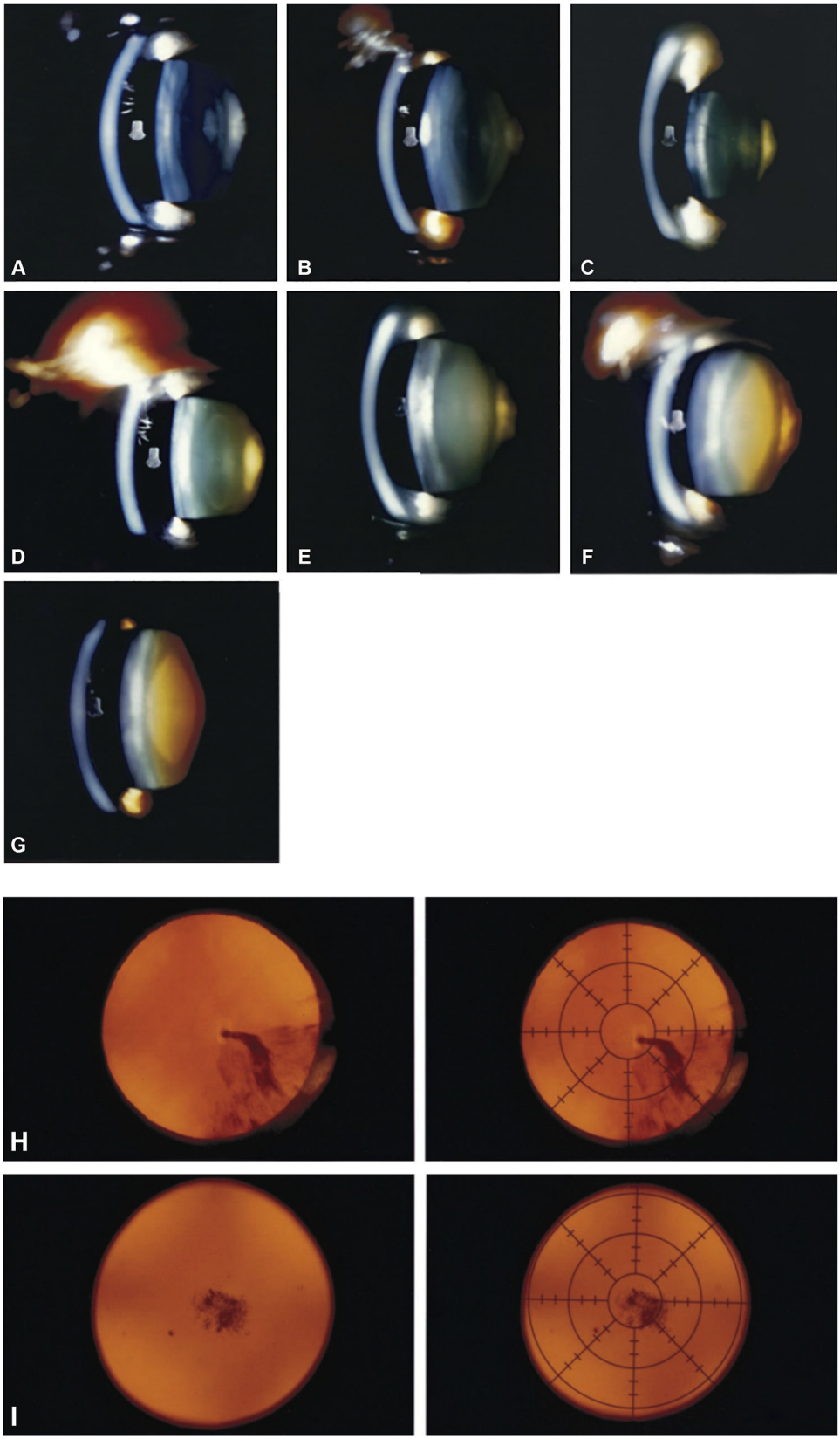

Figure 1. Reading center grading system for age-related cataract.

A-G, Nuclear cataract grading by comparison of 45-degree slit-lamp photograph with 7 standard photographs: 1 (no opacity) to 7 (extremely severe opacity). **A—G,** Standard photographs 1 through 7. **H,** Cortical cataract grading by percentage area involvement of the central 2 circles of the grid (5-mm diameter circular area) on retroillumination photograph. Left: Retroillumination photograph of cortical opacity. Right: Retroillumination photograph of cortical opacity with overlying grid. The cortical opacity occupies 22% of the central 2 circles of the grid. **I,** Posterior subcapsular cataract grading by percentage area involvement of the central 2 circles of the grid (5-mm diameter circular area) on retroillumination photograph. Left: Retroillumination photograph of posterior subcapsular opacity. Right: Retroillumination photograph of posterior subcapsular opacity with overlying grid. The posterior subcapsular opacity occupies 15% of the central 2 circles of the grid.

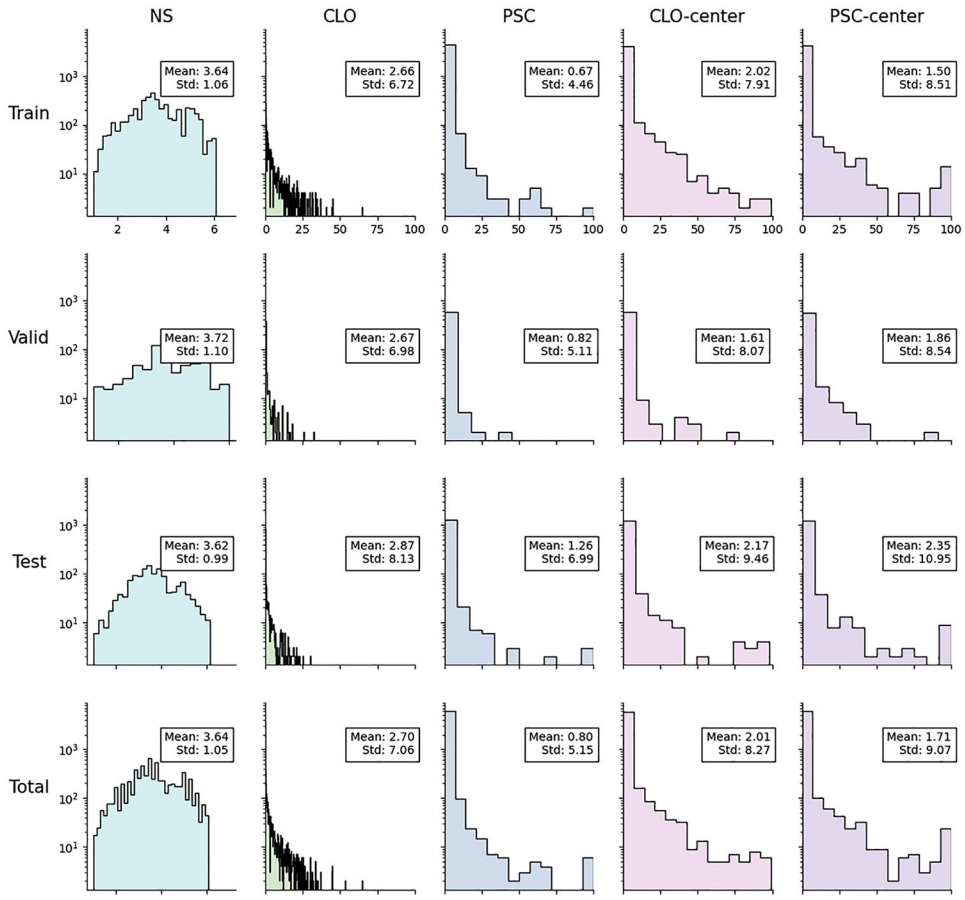

Figure 2.

Distributions of the cataract variables in the study population. The x-axis shows the grading scales, and the y-axis shows the associated frequencies on a logarithmic scale. CLO = cortical cataract; NS = nuclear sclerosis; PSC = posterior subcapsular cataract; Std = standard deviation.

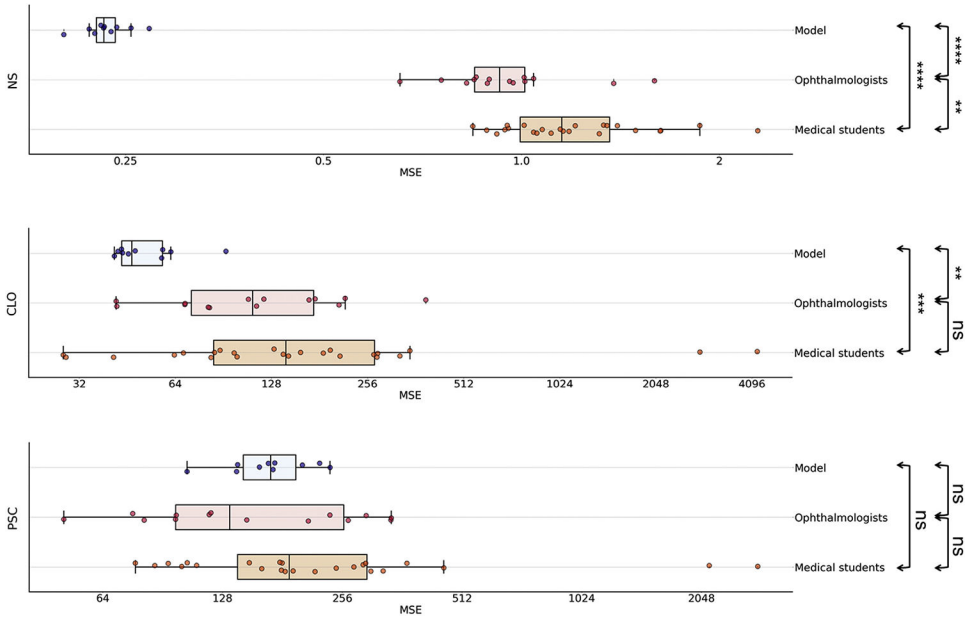

Figure 3.

Box plots showing the mean squared error (MSE) results on a logarithmic scale for the 10 deep learning models, the 14 ophthalmologists, and the 24 medical students for the 3 primary grading variables (NS, 0.9–7.1; CLO, 0%–100%; PSC, 0%–100%). The vertical line of the boxes represents the median MSE score, and the boxes represent the first and third quartiles. The whiskers represent quartile 1 – (1.5 × interquartile range) and quartile 3 + (1.5 × interquartile range). The dots represent the individual MSE result for each model or human grader. ****$P \leq 0.0001$; ***$P \leq 0.001$; **$P \leq 0.01$; ns, $P > 0.05$ (Mann–Whitney U test). CLO = cortical lens opacity; NS = nuclear sclerosis; PSC = posterior subcapsular cataract.

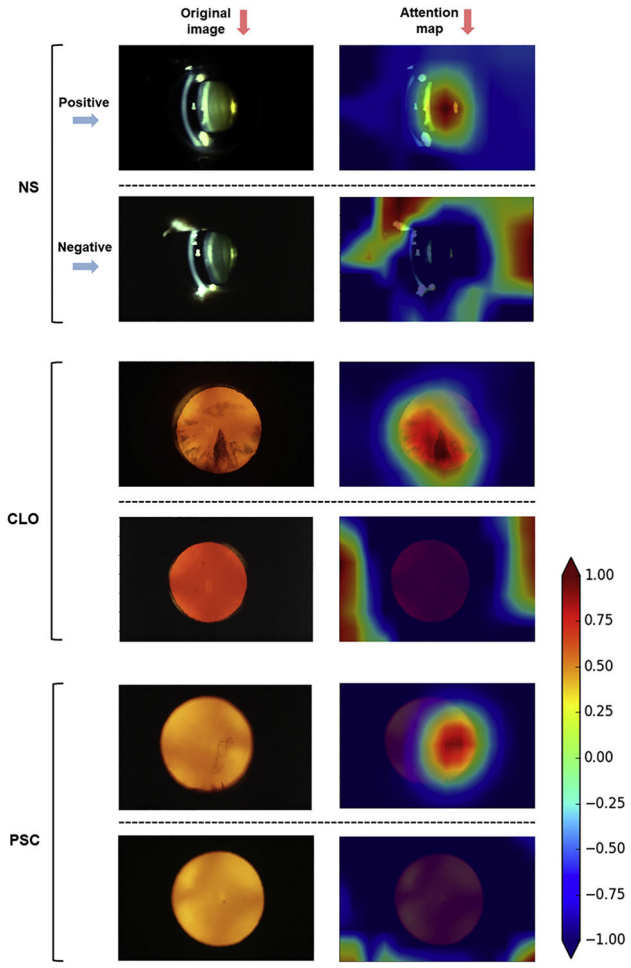

Figure 4.

Deep learning attention maps (right) overlaid on representative retroillumination images and 45-degree slit-lamp images (left). For each of the 3 cataract types (nuclear, cortical, and posterior subcapsular), 1 positive example (above) and 1 negative example (below) are shown (or more severe and less severe for nuclear cataract). For each image, the attention map demonstrates quantitatively the relative contributions made by each pixel to the grading prediction. The heatmap scale for the attention maps is also shown: signal range from −1.00 (purple) to +1.00 (brown). In the positive case of cortical cataract, the areas of high signal corresponded closely to the location and extent of the opacity. In the negative case, no areas of high signal were observed in the lens distribution. In the positive case of PSC, the area of high signal corresponded closely to the location and shape of the opacity (a single vertically elongated plaque). In the negative case, no areas of high signal were observed in the lens distribution. In the severe case of nuclear cataract, the area of high signal corresponded to the location of the lens nucleus. In the mild case, no areas of high signal were observed in the distribution of the lens nucleus. Nuclear sclerosis severe case: reading center grading of 5.3, automated prediction of 5.2. Nuclear sclerosis mild case: reading center grading of 2.5, automated prediction of 2.6. Cortical lens opacity positive case: reading center grading of 41.6%, automated prediction of 43.6%. Cortical lens opacity negative case: reading center grading of 0%, automated prediction of 0%. Posterior subcapsular cataract positive case:

reading center grading of 19.8%, automated prediction of 18.6%. Posterior subcapsular cataract negative case: reading center grading of 0%, automated prediction of 0%. CLO = cortical lens opacity; NS = nuclear sclerosis; PSC = posterior subcapsular cataract.

Table 1.

Characteristics of the Study Population

	Training Set	Validation Set	Test Set	Total
Participants (eyes)	403 (794)	57 (112)	116 (231)	576 (1137)
Female (%)	54.6	61.4	50.0	54.3
Mean age (yrs)	68.6	69.6	68.0	68.6
Image triplets	4425	598	1310	6333

Table 2.

Hyperparameter Values Used for Model Training

Hyperparameter	Value
Image size	501 × 334 pixels
Fully connected layers	1024, 128
Dropout ratio	0.5
Learning rate	0.0001
Batch size	16
Loss function	MSE

MSE = mean squared error.

Table 3.

Performance of the Deep Learning Models on the Full Test Set

	NS	CLO	PSC	CLO-Center	PSC-Center
MSE: mean (SD)	0.23 (0.01)	13.1 (1.6)	16.6 (2.4)	53.7 (4.9)	51.9 (6.5)

For each variable, 10 deep learning models were trained and tested separately.

CLO = cortical lens opacity; MSE = mean squared error; NS = nuclear sclerotic cataract; PSC = posterior subcapsular cataract; SD = standard deviation.

Table 4.

Performance Metrics of the Deep Learning Models versus the Human Graders on the Subset of the Test Set

MSE: Mean (SD)	NS	CLO	PSC	CLO-Center	PSC-Center
Deep learning models	0.23 (0.02)	53.5 (14.8)	171.9 (38.9)	252.5 (35.8)	520.1 (72.7)
Ophthalmologists (n = 14)	0.98 (0.24)	134.9 (89.9)	176.8 (98.0)	238.7 (125.8)	454.0 (207.1)
Medical students (n = 24)	1.24 (0.34)	433.6 (962.1)	398.2 (645.4)	588.5 (1149.6)	678.3 (640.2)
All human graders (n = 38)	1.14 (0.33)	323.5 (780.0)	316.6 (527.3)	459.7 (932.2)	595.7 (535.1)

CLO = cortical lens opacity; MSE = mean squared error; NS = nuclear sclerosis; PSC = posterior subcapsular cataract; SD = standard deviation.