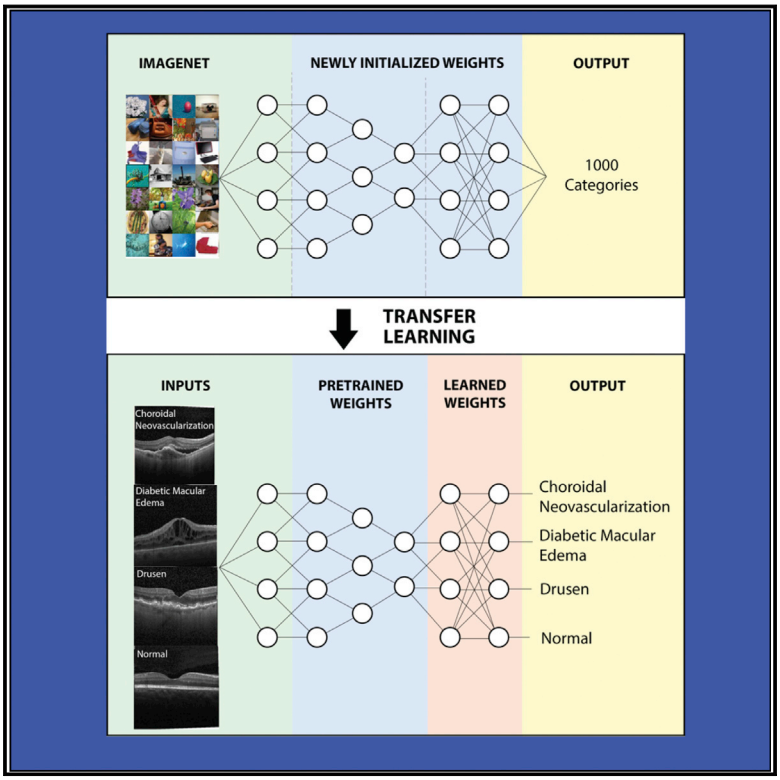
Resource

Cell

Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning

Graphical Abstract



Authors

Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, ..., M. Anthony Lewis, Huimin Xia, Kang Zhang

Correspondence

kang.zhang@gmail.com

In Brief

Image-based deep learning classifies macular degeneration and diabetic retinopathy using retinal optical coherence tomography images and has potential for generalized applications in biomedical image interpretation and medical decision making.

Highlights

- An artificial intelligence system using transfer learning techniques was developed
- It effectively classified images for macular degeneration and diabetic retinopathy
- It also accurately distinguished bacterial and viral pneumonia on chest X-rays
- This has potential for generalized high-impact application in biomedical imaging

Kermany et al., 2018, Cell *172*, 1122–1131
February 22, 2018
https://doi.org/10.1016/j.cell.2018.02.010

CellPress

Cell

Resource

Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning

Daniel S. Kermany,[1,2,14] Michael Goldbaum,[2,14] Wenjia Cai,[2,14] Carolina C.S. Valentim,[2,14] Huiying Liang,[1,14] Sally L. Baxter,[2,14] Alex McKeown,[3] Ge Yang,[2] Xiaokang Wu,[4] Fangbing Yan,[4] Justin Dong,[1] Made K. Prasadha,[2] Jacqueline Pei,[1,2] Magdalene Y.L. Ting,[2] Jie Zhu,[1,5] Christina Li,[2] Sierra Hewett,[1,2] Jason Dong,[1] Ian Ziyar,[2] Alexander Shi,[2] Runze Zhang,[2] Lianghong Zheng,[6] Rui Hou,[5] William Shi,[2] Xin Fu,[1,2] Yaou Duan,[2] Viet A.N. Huu,[1,2] Cindy Wen,[2] Edward D. Zhang,[1,2] Charlotte L. Zhang,[1,2] Oulan Li,[1,2] Xiaobo Wang,[7] Michael A. Singer,[8] Xiaodong Sun,[9] Jie Xu,[10] Ali Tafreshi,[3] M. Anthony Lewis,[11] Huimin Xia,[1] and Kang Zhang[1,2,4,12,13,15,*]

[1]Guangzhou Women and Children's Medical Center, Guangzhou Medical University, 510005 Guangzhou, China
[2]Shiley Eye Institute, Institute for Engineering in Medicine, Institute for Genomic Medicine, University of California, San Diego, La Jolla, CA 92093, USA
[3]Heidelberg Engineering, Heidelberg, Germany
[4]Molecular Medicine Research Center, State Key Laboratory of Biotherapy, The National Clinical Research Center of Senile Disease, West China Hospital, Sichuan University, Chengdu, China
[5]Guangzhou KangRui Biological Pharmaceutical Technology Company, 510005 Guangzhou, China
[6]YouHealth AI, 510005 Guangzhou, China
[7]Beihai Hospital, Dalian, 116021, China
[8]Department of Ophthalmology, University of Texas Health Science Center, San Antonio, TX 78229, USA
[9]Shanghai Key Laboratory of Ocular Fundus Diseases, Shanghai General Hospital, Shanghai JiaoTong University, 200080 Shanghai, China
[10]Beijing Instute of Ophthalmology, Beijing Tongren Eye Center, Beijing Tongren Hospital, Capital Medical University, Beijing, China
[11]Qualcomm, San Diego, CA 92121, USA
[12]Guangzhou Regenerative Medicine and Health Guangdong Laboratory, 510005 Guangzhou, China
[13]Veterans Administration Healthcare System, San Diego, CA 92037, USA
[14]These authors contributed equally
[15]Lead Contact
*Correspondence: kang.zhang@gmail.com
https://doi.org/10.1016/j.cell.2018.02.010

SUMMARY

The implementation of clinical-decision support algorithms for medical imaging faces challenges with reliability and interpretability. Here, we establish a diagnostic tool based on a deep-learning framework for the screening of patients with common treatable blinding retinal diseases. Our framework utilizes transfer learning, which trains a neural network with a fraction of the data of conventional approaches. Applying this approach to a dataset of optical coherence tomography images, we demonstrate performance comparable to that of human experts in classifying age-related macular degeneration and diabetic macular edema. We also provide a more transparent and interpretable diagnosis by highlighting the regions recognized by the neural network. We further demonstrate the general applicability of our AI system for diagnosis of pediatric pneumonia using chest X-ray images. This tool may ultimately aid in expediting the diagnosis and referral of these treatable conditions, thereby facilitating earlier treatment, resulting in improved clinical outcomes.

INTRODUCTION

Artificial intelligence (AI) has the potential to revolutionize disease diagnosis and management by performing classification difficult for human experts and by rapidly reviewing immense amounts of images. Despite its potential, clinical interpretability and feasible preparation of AI remains challenging.

The traditional algorithmic approach to image analysis for classification previously relied on (1) handcrafted object segmentation, followed by (2) identification of each segmented object using statistical classifiers or shallow neural computational machine-learning classifiers designed specifically for each class of objects, and finally (3) classification of the image (Goldbaum et al., 1996). Creating and refining multiple classifiers required many skilled people and much time and was computationally expensive (Chaudhuri et al., 1989; Hoover and Goldbaum, 2003; Hoover et al., 2000).

The development of convolutional neural network layers has allowed for significant gains in the ability to classify images and detect objects in a picture (Krizhevsky et al., 2017; Zeiler and Fergus, 2014). These are multiple processing layers to which image analysis filters, or convolutions, are applied. The abstracted representation of images within each layer is constructed by systematically convolving multiple filters across the image, producing a feature map that is used as input to the following layer. This architecture makes it possible to process images in the form of pixels as input and to give the desired

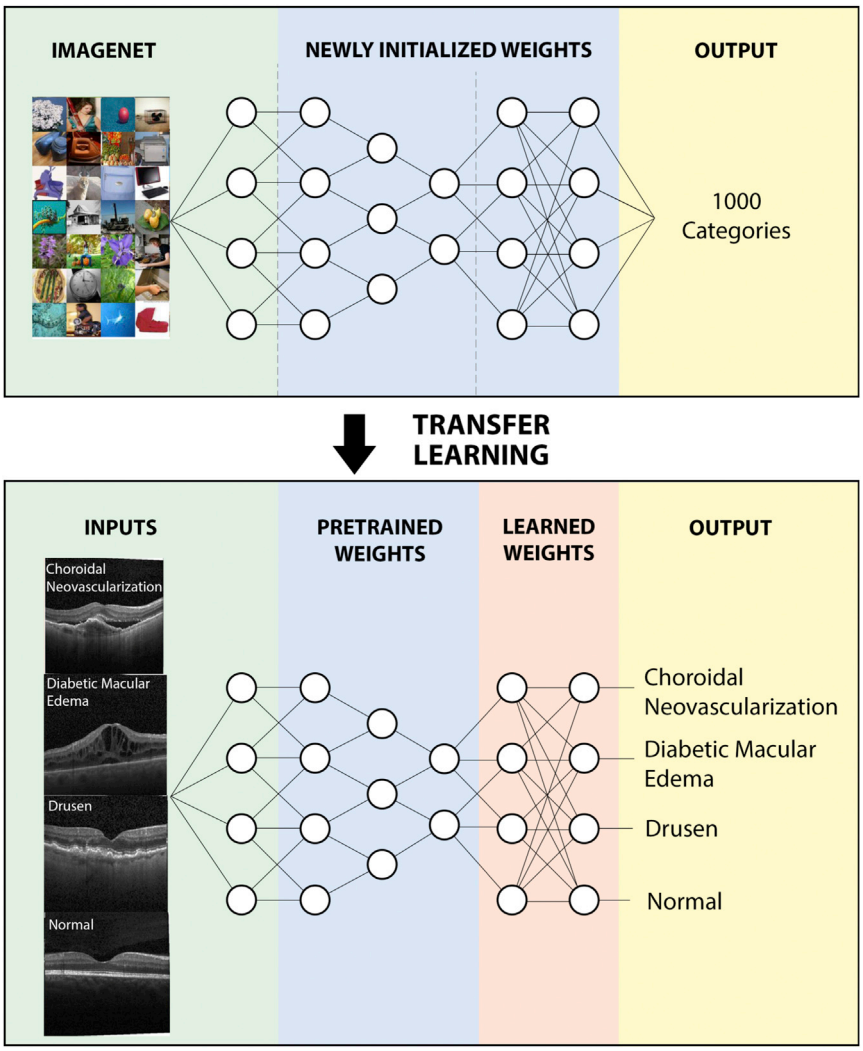


Figure 1. Schematic of a Convolutional Neural Network

Schematic depicting how a convolutional neural network trained on the ImageNet dataset of 1,000 categories can be adapted to significantly increase the accuracy and shorten the training duration of a network trained on a novel dataset of OCT images. The locally connected (convolutional) layers are frozen and transferred into a new network, while the final, fully connected layers are recreated and retrained from random initialization on top of the transferred layers.

classification as output. The image-to-classification approach in one classifier replaces the multiple steps of previous image analysis methods.

One method of addressing a lack of data in a given domain is to leverage data from a similar domain, a technique known as transfer learning. Transfer learning has proven to be a highly effective technique, particularly when faced with domains with limited data (Donahue et al., 2013; Razavian et al., 2014; Yosinski et al., 2014). Rather than training a completely blank network, by using a feed-forward approach to fix the weights in the lower levels already optimized to recognize the structures found in images in general and retraining the weights of the upper levels with back propagation, the model can recognize the distinguishing features of a specific category of images, such as images of the eye, much faster and with significantly fewer training examples and less computational power (Figure 1).

In this study, we sought to develop an effective transfer learning algorithm to process medical images to provide an accurate and timely diagnosis of key pathology in each image. The primary illustration of this technique involved optical coherence tomography (OCT) images of the retina, but the algorithm was also tested in a cohort of pediatric chest radiographs to validate the generalizability of this technique across multiple imaging modalities.

RESULTS

The primary application of our transfer learning algorithm was in the diagnosis of retinal OCT images. Spectral-domain OCT uses light to capture high-resolution *in vivo* optical cross sections of the retina that can be assembled into three-dimensional-volume images of living retinal tissue. It has become one of the most commonly performed medical imaging procedures, with approximately 30 million OCT scans performed each year worldwide (Swanson and Fujimoto, 2017). OCT imaging is now a standard of care for guiding the diagnosis and treatment of some of the leading causes of blindness worldwide: age-related macular degeneration (AMD) and diabetic macular edema. Almost 10 million individuals suffer from AMD in the United States, and each year, more than 200,000 people develop choroidal neovascularization, a severe blinding form of advanced AMD (Ferrara, 2010; Friedman et al., 2004; Wong et al., 2014). In addition, nearly 750,000 individuals aged 40 or older suffer from diabetic macular edema (Varma et al., 2014), a vision-threatening form of diabetic retinopathy that involves the accumulation of fluid in the central retina. The prevalence of these diseases will likely increase even further over time due to the aging population and the global diabetes epidemic. Fortunately, the advent and widespread utilization of anti-vascular endothelial growth factor (anti-VEGF) medications has revolutionized the treatment of exudative retinal diseases (Kaiser et al., 2007; Ferrara, 2010), allowing patients to retain useful vision and quality of life. OCT is critical to guiding the administration of anti-VEGF therapy by providing a clear cross-sectional representation of the retinal pathology in these conditions (Figure 2A), allowing visualization of individual retinal layers, which is impossible with clinical examination by the human eye or by color fundus photography.

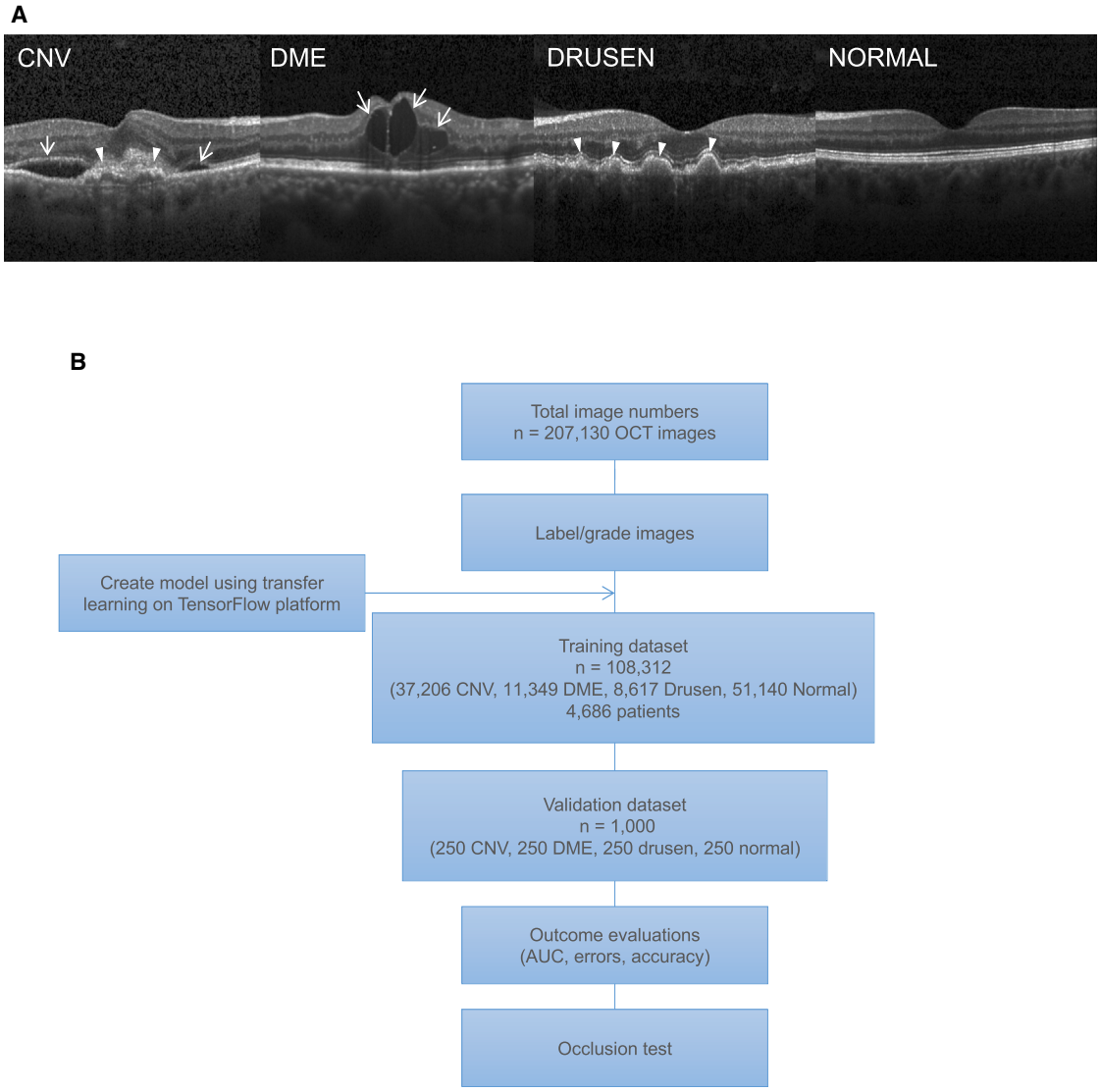


Figure 2. Representative Optical Coherence Tomography Images and the Workflow Diagram

(A) (Far left) choroidal neovascularization (CNV) with neovascular membrane (white arrowheads) and associated subretinal fluid (arrows). (Middle left) Diabetic macular edema (DME) with retinal-thickening-associated intraretinal fluid (arrows). (Middle right) Multiple drusen (arrowheads) present in early AMD. (Far right) Normal retina with preserved foveal contour and absence of any retinal fluid/edema.

(B) Workflow diagram showing overall experimental design describing the flow of optical coherence tomography (OCT) images through the labeling and grading process followed by creation of the transfer learning model, which then underwent training and subsequent testing. The training dataset only included images that passed sufficient quality and diagnostic standards from the initial collected dataset.

See also Table S1.

Patient and Image Characteristics

We initially obtained 207,130 OCT images. 108,312 images (37,206 with choroidal neovascularization, 11,349 with diabetic macular edema, 8,617 with drusen, and 51,140 normal) from 4,686 patients passed initial image quality review and were used to train the AI system. The model was tested with 1,000 images (250 from each category) from 633 patients. Patient characteristics for each diagnosis category are listed in Table S1. After 100 epochs (iterations through the entire dataset), the training was stopped due to the absence of further improvement in both accuracy (Figure 3A) and cross-entropy loss (Figure 3B).

Performance of the Model

We evaluated our AI system in diagnosing the most common blinding retinal diseases. This AI system categorized images with choroidal neovascularization and images with diabetic macular edema as "urgent referrals." These conditions would demand relatively urgent referral to an ophthalmologist for definitive anti-VEGF treatment; if treatment is delayed, there is

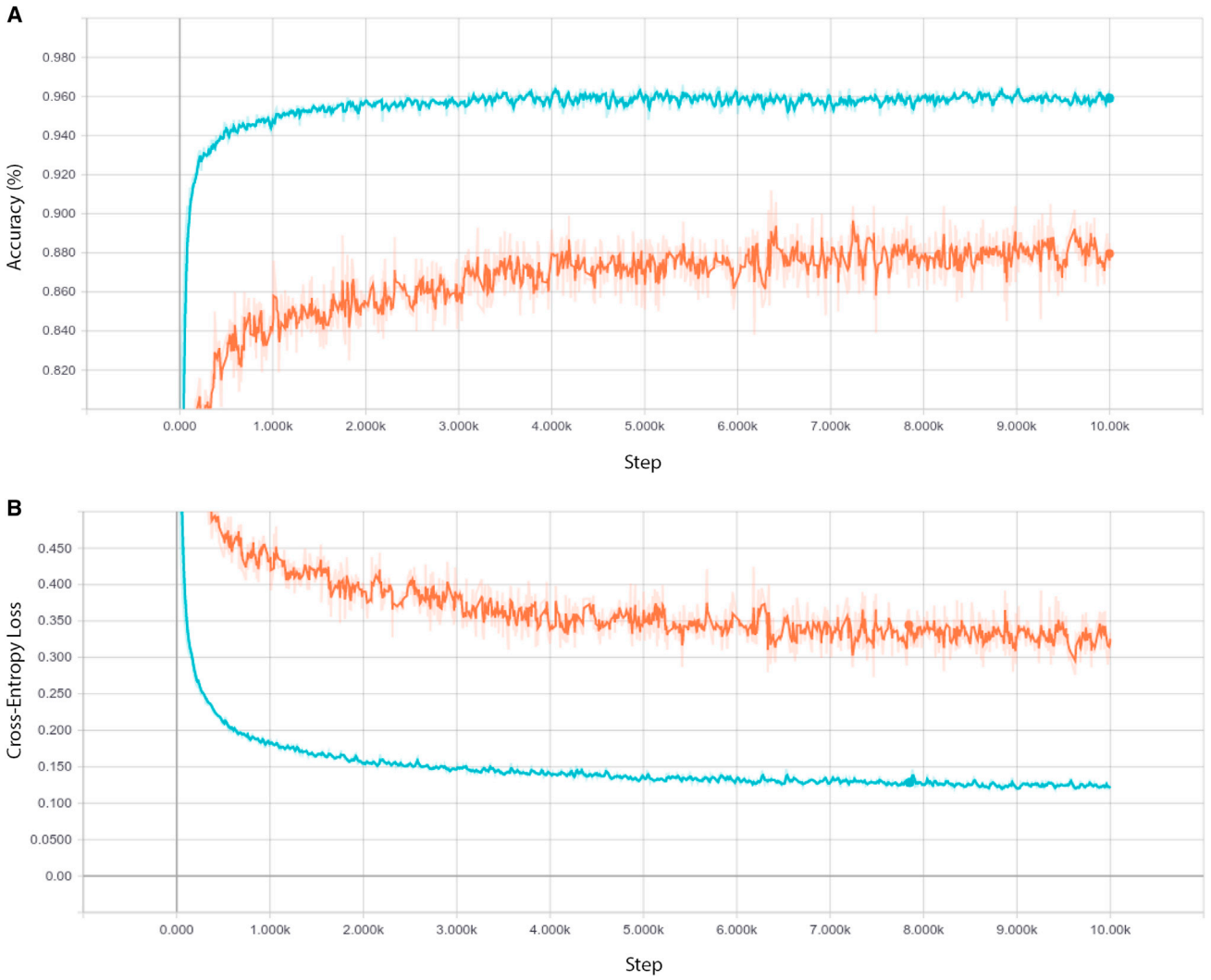


Figure 3. Plot Showing Performance in the Training and Validation Datasets Using TensorBoard

Accuracy is plotted against the training step (A), and cross-entropy loss is plotted against the training step (B) during the length of the training of the multi-class classifier over the course of 10,000 steps. Plots were normalized with a smoothing factor of 0.6 to clearly visualize trends. The validation accuracy and loss show better performance, since images with more noise and lower quality were also included in the training set to reduce overfitting and help generalization of the classifier. Training dataset: orange. Validation dataset: blue.

See also Figure S1.

increased risk of bleeding, scarring, or other downstream complications that cause irreversible vision impairment. The system categorized images with drusen, which are lipid deposits present in the dry form of macular degeneration, as "routine referrals." Anti-VEGF medications are not indicated for dry macular degeneration; therefore, referral to an eye specialist for drusen is less urgent. Normal images were labeled for "observation." In a multi-class comparison between choroidal neovascularization, diabetic macular edema, drusen, and normal, we achieved an accuracy of 96.6% (Figure 4), with a sensitivity of 97.8%, a specificity of 97.4%, and a weighted error of 6.6%. Receiver operating characteristic (ROC) curves were generated to evaluate the model's ability to distinguish urgent referrals (defined as choroidal neovascularization or diabetic macular edema) from drusen and normal exams. The area under the ROC curve was 99.9% (Figure 4).

We also trained a "limited model" classifying between the same four categories but only using 1,000 images randomly selected from each class during training to compare transfer learning performance using limited data compared to results using a large dataset. Using the same testing images, the model achieved an accuracy of 93.4%, with a sensitivity of 96.6%, a specificity of 94.0%, and a weighted error of 12.7%. The ROC curves distinguishing urgent referrals (i.e., distinguishing images with choroidal neovascularization or diabetic macular edema from normal images had an area under the curve of 98.8%.

Binary classifiers were also implemented to compare choroidal neovascularization/diabetic macular edema/drusen from normal

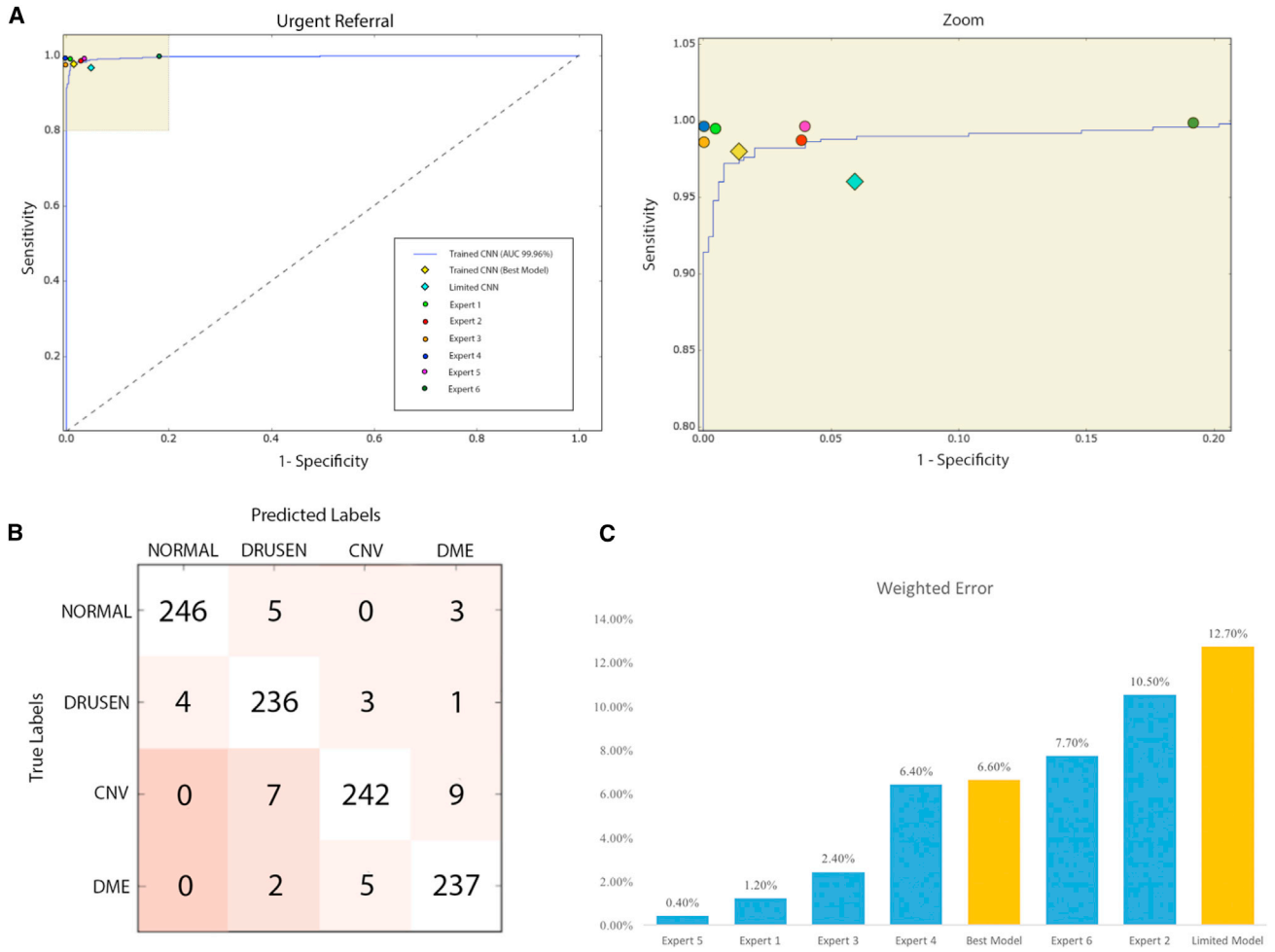


Figure 4. Multi-class Comparison between Choroidal Neovascularization, Diabetic Macular Edema, Drusen, and Normal

(A) Receiver operating characteristic (ROC) curve for "urgent referrals" (CNV and DME detection) with human expert performance for comparison. The area under the ROC curve was 99.9%. The zoomed area shows that the most accurate model demonstrates a performance that rivals that of six human experts.

(B) Confusion table of best model's classification of the validation image set. The model successfully scored all urgent referrals as higher than observation.

(C) Weighted error results based on penalties in Figure S4 depicting neural networks in gold and human experts in blue.

See also Figures S2, S3, and S4 and Table S2.

using the same datasets in order to determine a breakdown of the model's performance (Figure S1). The classifier distinguishing choroidal neovascularization images from normal images achieved an accuracy of 100.0%, with a sensitivity of 100.0% and specificity of 100.0%. The area under the ROC curve was 100.0% (Figure S2A). The classifier distinguishing diabetic macular edema images from normal images achieved an accuracy of 98.2%, with a sensitivity of 96.8% and specificity of 99.6%. The area under the ROC curve was 99.87% (Figure S2B). The classifier distinguishing drusen images from normal images achieved an accuracy of 99.0%, with a sensitivity of 98.0% and specificity of 99.2%. The area under the ROC curve was 99.96% (Figure S2C).

Comparison of the Model with Human Experts

An independent test set of 1,000 images from 633 patients was used to compare the AI network's referral decisions with the decisions made by human experts. Six experts with significant clinical experience in an academic ophthalmology center were instructed to make a referral decision on each test patient using only the patient's OCT images. Performance on the clinically most important decision of distinguishing patients needing urgent referral (those with choroidal neovascularization or diabetic macular edema) compared to normal patients is displayed as a ROC curve, and this performance was comparable between the AI system and the human experts (Figure 4A).

Having established a standard expert performance evaluation system, we next compared the potential impact of patient referral decisions between our network and human experts. The sensitivities and specificities of the experts were plotted on the ROC curve of the trained model, and the differences in diagnostic performance, measured by likelihood ratios, between the model and the human experts were determined to be statistically similar within a 95% confidence interval (Figure S3).

However, the pure error rate does not accurately reflect the impact that a wrong referral decision might have on the outcome of an individual patient. To illustrate, a false-positive result occurs when a patient is normal or has drusen but is inaccurately labeled as an urgent referral, and this can cause undue distress or unnecessary investigation for the patient and place extra burdens on the healthcare system. However, a false-negative result is far more serious, because in this instance, a patient with choroidal neovascularization or diabetic macular edema is not appropriately referred, which could result in irreversible visual loss. To account for these issues, weighted error scoring was incorporated during model evaluation and expert testing (Figure S4A). By assigning these penalty points to each decision made by the model and the experts, we computed the average error of each.

The best convolutional neural network model yielded a score of 6.6% under this weighted error system. The weighted error of the experts ranged from 0.4% to 10.5%, with a mean weighted error of 4.8% (Table S2). The exact breakdown of each expert's performance regarding the correlation of their predicted labels with the true labels is depicted as confusion matrices in Figure S4B. As seen in Figure 4, the best model outperformed some human experts based on this weighted scale and on the ROC curve.

Occlusion Testing

We performed an occlusion test on 491 images to identify the areas contributing most to the neural network's assignment of the predicted diagnosis. This testing successfully identified the region of interest in 94.7% of images that contributed the highest importance to the deep-learning algorithm (Figure 5A; see also Figure S5 for additional examples). Drusen were located correctly through occlusion testing in 100% of all the images, while choroidal neovascularization yielded an accuracy of 94.0% and diabetic macular edema yielded an accuracy of 91.0% (Table S3). Furthermore, these regions identified by occlusion testing were also verified by human experts to be the most clinically significant areas of pathology.

Application of the AI System for Pneumonia Detection Using Chest X-Ray Images

To investigate the generalizability of our AI system in the diagnosis of common diseases, we applied the same transfer learning framework to the diagnosis of pediatric pneumonia. According to the World Health Organization (WHO), pneumonia kills about 2 million children under 5 years old every year and is consistently estimated as the single leading cause of childhood mortality (Rudan et al., 2008), killing more children than HIV/AIDS, malaria, and measles combined (Adegbola, 2012). The WHO reports that nearly all cases (95%) of new-onset childhood clinical pneumonia occur in developing countries, particularly in Southeast Asia and Africa. Bacterial and viral pathogens are the two leading causes of pneumonia (Mcluckie, 2009) but require very different forms of management. Bacterial pneumonia requires urgent referral for immediate antibiotic treatment, while viral pneumonia is treated with supportive care. Therefore, accurate and timely diagnosis is imperative. One key element of diagnosis is radiographic data, since chest X-rays are routinely obtained as standard of care and can help differentiate between different types of pneumonia (Figure S6). However, rapid radiologic interpretation of images is not always available, particularly in the low-resource settings where childhood pneumonia has the highest incidence and highest rates of mortality. To this end, we also investigated the effectiveness of our transfer learning framework in classifying pediatric chest X-rays to detect pneumonia and furthermore to distinguish viral and bacterial pneumonia to facilitate rapid referrals for children needing urgent intervention.

We collected and labeled a total of 5,232 chest X-ray images from children, including 3,883 characterized as depicting pneumonia (2,538 bacterial and 1,345 viral) and 1,349 normal, from a total of 5,856 patients to train the AI system. The model was then tested with 234 normal images and 390 pneumonia images (242 bacterial and 148 viral) from 624 patients. After 100 epochs (iterations through the entire dataset) of the model, the training was stopped due to the absence of further improvement in both loss and accuracy (Figures 6A and 6B).

In the comparison of chest X-rays presenting as pneumonia versus normal, we achieved an accuracy of 92.8%, with a sensitivity of 93.2% and a specificity of 90.1%. The area under the ROC curve for detection of pneumonia from normal was 96.8% (Figure 6E). Binary comparison of bacterial and viral pneumonia resulted in a test accuracy of 90.7%, with a sensitivity of 88.6% and a specificity of 90.9% (Figures 6C and 6D). The area under the ROC curve for distinguishing bacterial and viral pneumonia was 94.0% (Figure 6F).

DISCUSSION

In this study, we describe a general AI platform for the diagnosis and referral of two common causes of severe vision loss: diabetic macular edema and choroidal neovascularization seen in neovascular AMD. By employing a transfer learning algorithm, our model demonstrated competitive performance of OCT image analysis without the need for a highly specialized deep-learning machine and without a database of millions of example images (STAR Methods). Moreover, the model's performance in diagnosing retinal OCT images was comparable to that of human experts with significant clinical experience with retinal diseases. When the model was trained with a much smaller number of images (about 1,000 from each class), it retained high performance in accuracy, sensitivity, specificity, and area under the ROC curve for achieving the correct diagnosis and referral, thereby illustrating the power of the transfer learning system to make highly effective classifications, even with a very limited training dataset.

Although our AI platform was trained and validated using the Heidelberg Spectralis imaging system, the Digital Imaging and Communications in Medicine (DICOM) standards make the OCT images from different manufacturers (e.g., Zeiss and Optovue) reasonably consistent. The goal of this preliminary approach was to develop a system and demonstrate the soundness of the methods. Future studies could entail the use of images from different manufacturers in both the training and testing datasets so that the system will be universally useful. Moreover, the efficacy of the transfer learning technique for image analysis very likely extends beyond the realm of OCT images

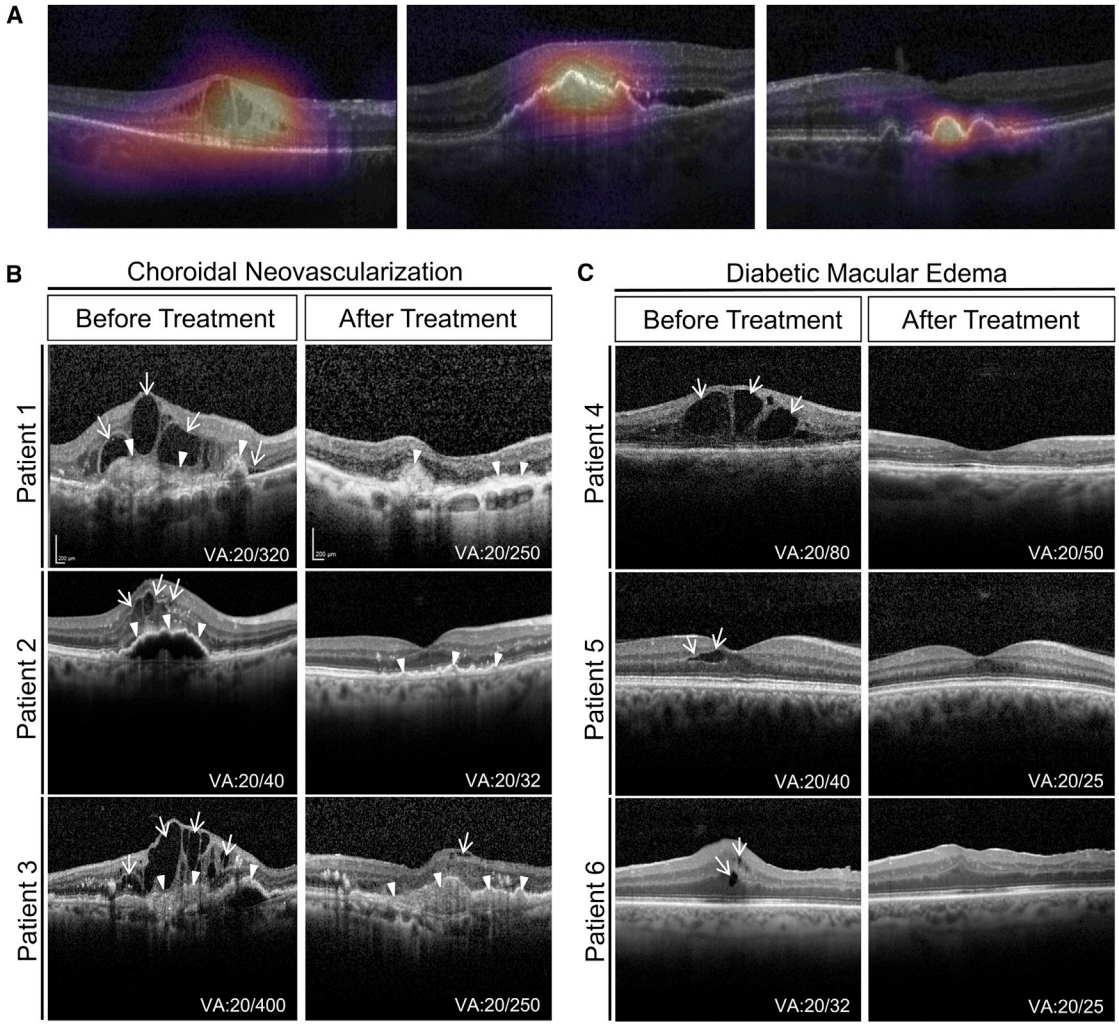


Figure 5. Occlusion Maps and Longitudinal Follow-up OCT Images Comparing Retinal Structural Changes before and after Anti-VEGF Therapy

(A) Occlusion maps highlighting areas of pathology in diabetic macular edema (left), choroidal neovascularization (middle), and drusen (right). An occlusion map was generated by convolving an occluding kernel across the input image. The occlusion map is created after prediction by assigning the softmax probability of the correct label to each occluded area. The occlusion map can then be superimposed on the input image to highlight the areas the model considered important in making its diagnosis.

(B and C) Horizontal cross-section OCT images through the fovea of patients with wet AMD (B) or diabetic retinopathy with macular edema (C) before and after three monthly intravitreal injections of bevacizumab. Both intraretinal and subretinal fluid (white arrows) lessened after treatment. Scar tissue of choroidal neovascularization remained (arrow heads). All visual accurity (VA) was improved: 20/320 to 20/250, 5 months (patient 1); 20/40 to 20/32, 9 months (patient 2); 20/400 to 20/250, 3 months (patient 3); 20/80 to 20/50, 7 months (patient 4); 20/40 to 20/25; 7 months (patient 5); and 20/32 to 20/25, 7 months (patient 6). See also Figure S5 and Table S3.

and ophthalmology—in principle, the techniques we have described here could potentially be employed in a wide range of medical images across multiple disciplines, and in fact, we provide a direct illustration of its wide applicability by demonstrating its efficacy in analysis of chest X-ray images.

Occlusion testing was performed to identify the areas of greatest importance used by the model in assigning a diagnosis. The greatest benefit of an occlusion test is that it reveals insights into the decisions of neural networks, which are infamously known as ''black boxes'' with no transparency. Since this test was performed after training was completed, it demystified the algorithm without affecting its results. The occlusion test also confirmed that the network made its decisions using accurate distinguishing features, which can be shared with a healthcare professional. All areas containing drusen were recognized correctly on all images used for testing, while the diabetic macular edema and choroidal neovascularization occlusion tests occasionally did not present a clear point of interest. This is likely due to the lesions and fluid pockets of choroidal neovascularization and diabetic macular edema sometimes presenting much larger than the occlusion window, while drusen tend to be smaller in size.

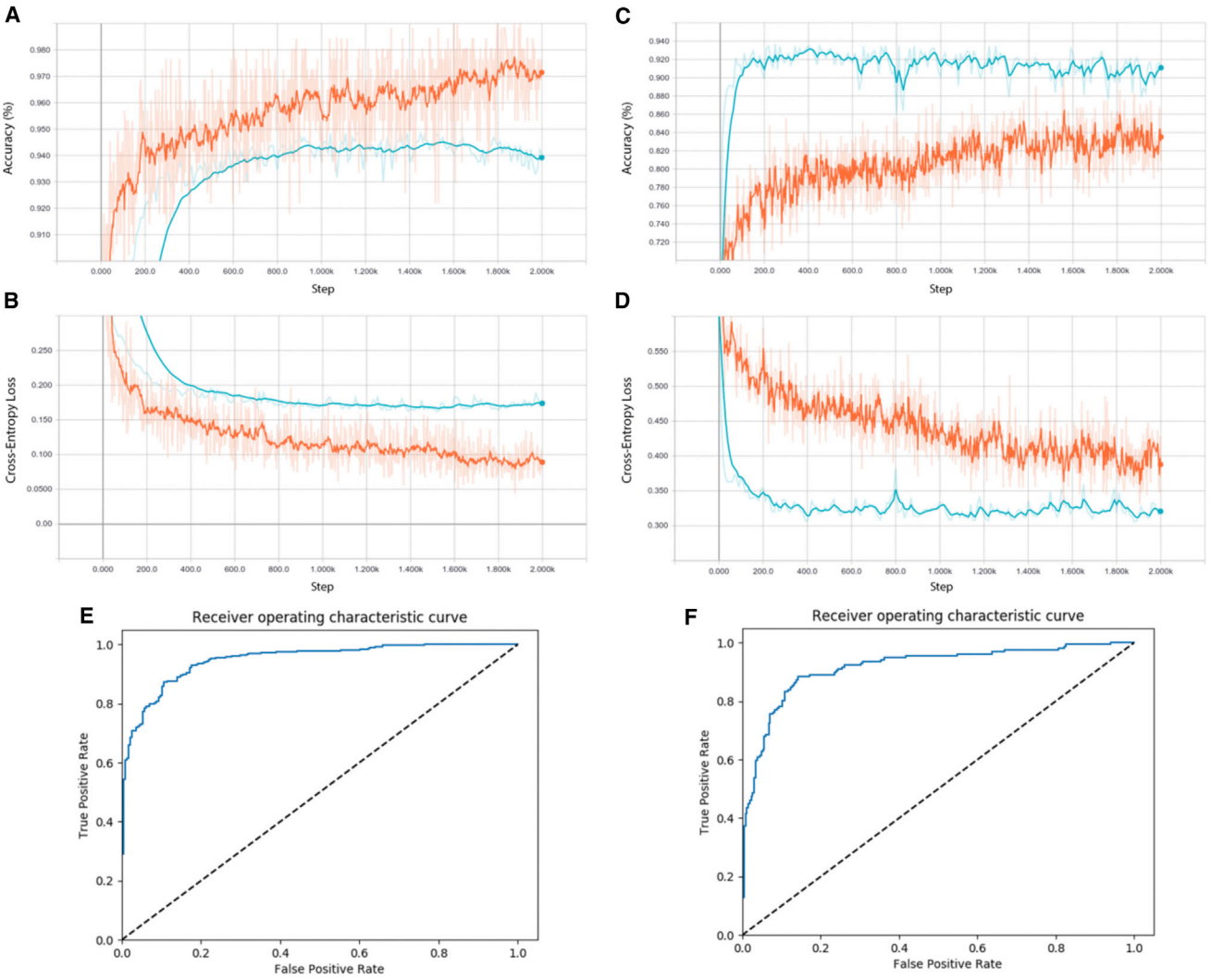


Figure 6. Plots Depicting Performance of Pneumonia Diagnosis using Chest X-Ray Images in the Training and Validation Datasets Using TensorBoard

(A–F) Comparisons were made for pneumonia versus normal (A) with cross-entropy loss plotted against the training step (B), as well as comparisons between bacterial pneumonia and viral pneumonia (C) and the associated cross-entropy loss (D). Plots were normalized with a smoothing factor of 0.6 in order to clearly visualize trends. The area under the ROC curve for detecting pneumonia versus normal was 96.8% (E). The area under the ROC curve for detecting bacterial versus viral pneumonia was 94.0% (F). Training dataset: orange. Validation dataset: blue.

See also Figure S6.

Although transfer learning allows the training of a highly accurate model with a relatively small training dataset, its performance would be inferior to that of a model trained from a random initialization on an extremely large dataset of OCT images, since even the internal weights can be directly optimized for OCT feature detection. However, in practice, a new convolutional neural network trained from random initialization, even with an unlimited supply of training data, would require weeks to achieve a good accuracy, whereas the multi-class holdout model implemented using transfer learning finished training and testing on different data in under 2 hr. Each binary classification and the limited model converged to a high accuracy in under 30 min. Since medical images are difficult to collect in the large amounts necessary to train a blank convolutional neural network, transfer learning using a pre-trained model trained on millions of various medical images would likely yield a more accurate model in much less time when retraining layers for other medical classifications.

The performance of our model depends highly on the weights of the pre-trained model. Therefore, the performance of this model would likely be enhanced when tested on a larger ImageNet dataset with more advanced deep-learning techniques and architecture. Further, the rapid progression and development of the field of convolutional neural networks applied outside of medical imaging would also improve the performance of our approach.

Finally, as mentioned earlier, we use OCT imaging as a demonstration of a generalized approach in medical image interpretation and subsequent decision making. Our framework effectively identified potential pathology on a tissue map to make a referral decision with performance comparable to (and sometimes even better than) human experts, enabling timely diagnosis of the two most common causes of irreversible severe vision loss. OCT is particularly useful in the management of retinal diseases because it has become critical to guiding anti-VEGF treatment for the intraretinal and/or subretinal fluid seen in many retinal conditions. This fluid often cannot be clearly visualized by the examiner's eyes or by color fundus photography. In addition, the OCT appearance often correlates well with visual acuity. The presence of fluid is typically associated with worse visual acuity, which improves once the fluid is resolved with anti-VEGF treatment (Figure 5B). As a testament to the value of this imaging modality, treatment decisions for exudative retinal diseases are now guided by OCT rather than by clinical examination or fundus photography, making this demonstration of AI-guided classification of images more clinically relevant than prior studies that have analyzed retinal fundus photographs, such as that from Gulshan et al. (2016). Given that OCT imaging has played such a crucial role in guiding treatment, extending the application of AI beyond diagnosis or classification of images and into the realm of making treatment recommendations is a promising area of future investigation.

Furthermore, our network represents a generalized platform that can potentially be applied to a wide range of medical imaging techniques (e.g., chest X-ray, MRI, computed tomography) to make a clinical diagnostic decision. We demonstrated this point by training our network on a dataset of chest X-ray images of pediatric pneumonia. Chest X-rays present a difficult classification task due to the relatively large amount of variable objects, specifically the imaged areas outside the lungs that are irrelevant to the diagnosis of pneumonia. The resulting high-accuracy model suggests that this AI system has the potential to effectively learn from increasingly complicated images with a high degree of generalization using a relatively small repository of data. By demonstrating efficacy with multiple imaging modalities and with a wide range of pathology, this transfer learning framework presents a compelling system for further exploration and analysis in biomedical imaging and more generalized application to an automated community-based AI system for the diagnosis and triage of common human diseases. By providing our data and codes in a publicly available database, we also hope that other biomedical researchers may use our work as a resource to improve the performance of future models and help drive the field forward. This could facilitate screening programs and create more efficient referral systems in all of medicine, particularly in remote or low-resource areas, leading to a broad clinical and public health impact.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Images from Human Subjects
- METHOD DETAILS
  - Image Labeling
  - Transfer Learning Methods
  - Expert Comparisons
  - Occlusion Test
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes six figures and three tables and can be found with this article online at https://doi.org/10.1016/j.cell.2018.02.010.

A video abstract is available at http://dx.doi.org/10.1016/j.cell.2018.02.010#mmc2.

ACKNOWLEDGMENTS

This study was funded by the National Key Research and Development Program of China (2017YFC1104600), National Natural Science Foundation of China (81771629 and 81700882), Guangzhou Women and Children's Medical Center, Guangzhou Regenerative Medicine and Health Guangdong Laboratory, the Richard Annesser Fund, the Michael Martin Fund, and the Dick and Carol Hertzberg Fund.

AUTHOR CONTRIBUTIONS

D.S.K., M.A.L., W.C., Justin Dong, C.C.S.V., G.Y., H.L., A.M., X. Wu, F.Y., J.Z., S.L.B., M.K.P., J.P., A.S., M.Y.L.T., C.L., S.H., Jason Dong, R.Z., L.Z., R.H., W.S., X.F., Y.D., V.A.N.H., I.Z., C.W., X. Wang, E.D.Z., C.L.Z., O.L., J.X., A.T., X.S., M.A.S., and H.X. collected and analyzed the data. K.Z. conceived the project. K.Z., D.S.K., M.G., and S.L.B. wrote the manuscript. All authors discussed the results and reviewed the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 1, 2017
Revised: December 31, 2017
Accepted: February 1, 2018
Published: February 22, 2018

REFERENCES

Adegbola, R.A. (2012). Childhood pneumonia as a global health priority and the strategic interest of the Bill & Melinda Gates Foundation. Clin. Infect. Dis. *54* (*Suppl 2*), S89–S92.

Chaudhuri, S., Chatterjee, S., Katz, N., Nelson, M., and Goldbaum, M. (1989). Detection of blood vessels in retinal images using two-dimensional matched filters. IEEE Trans. Med. Imaging *8*, 263–269.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. (2013). DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. Proceedings of the 31st International Conference on Machine Learning *32*, 647–655.

Ferrara, N. (2010). Vascular endothelial growth factor and age-related macular degeneration: from basic science to therapy. Nat. Med. *16*, 1107–1111.

Friedman, D.S., O'Colmain, B.J., Muñoz, B., Tomany, S.C., McCarty, C., de Jong, P.T., Nemesure, B., Mitchell, P., and Kempen, J.; Eye Diseases Prevalence Research Group (2004). Prevalence of age-related macular degeneration in the United States. Arch. Ophthalmol. *122*, 564–572.

Goldbaum, M., Moezzi, S., Taylor, A., Chatterjee, S., Boyd, J., Hunter, E., and Jain, R. (1996). Automated diagnosis and image understanding with object

extraction, object classification, and inferencing in retinal images. Proceedings of 3rd IEEE International Conference on Image Processing *3*, 695–698.

Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA *316*, 2402–2410.

Hoover, A., and Goldbaum, M. (2003). Locating the optic nerve in a retinal image using the fuzzy convergence of the blood vessels. IEEE Trans. Med. Imaging *22*, 951–958.

Hoover, A., Kouznetsova, V., and Goldbaum, M. (2000). Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. IEEE Trans. Med. Imaging *19*, 203–210.

Kaiser, P.K., Brown, D.M., Zhang, K., Hudson, H.L., Holz, F.G., Shapiro, H., Schneider, S., and Acharya, N.R. (2007). Ranibizumab for predominantly classic neovascular age-related macular degeneration: subgroup analysis of first-year ANCHOR results. Am. J. Ophthalmol. *144*, 850–857.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2017). ImageNet classification with deep convolutional neural networks. Commun. ACM *60*, 84–90.

Lee, C.S., Baughman, D.M., and Lee, A.Y. (2016). Deep Learning Is Effective for the Classification of OCT Images of Normal versus Age-Related Macular Degeneration. Ophthamol. Retina *1*, 322–327.

Mcluckie, A. (2009). Respiratory disease and its management, *Volume 57* (Springer).

Razavian, A.S., Azizpour, H., Sullivan, J., and Carlsson, S. (2014). CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 512–519.

Rudan, I., Boschi-Pinto, C., Biloglav, Z., Mulholland, K., and Campbell, H. (2008). Epidemiology and etiology of childhood pneumonia. Bull. World Health Organ. *86*, 408–416.

Swanson, E.A., and Fujimoto, J.G. (2017). The ecosystem that powered the translation of OCT from fundamental research to clinical and commercial impact [Invited]. Biomed. Opt. Express *8*, 1638–1664.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. In 2016 IWWW Conference on Computer Vision and Pattern Recognition, pp. 2818–2826.

Varma, R., Bressler, N.M., Doan, Q.V., Gleeson, M., Danese, M., Bower, J.K., Selvin, E., Dolan, C., Fine, J., Colman, S., and Turpcu, A. (2014). Prevalence of and risk factors for diabetic macular edema in the United States. JAMA Ophthalmol. *132*, 1334–1340.

Wong, W.L., Su, X., Li, X., Cheung, C.M., Klein, R., Cheng, C.Y., and Wong, T.Y. (2014). Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. Lancet Glob. Health *2*, e106–e116.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems *2*, 3320–3328.

Zeiler, M.D., and Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. Lect. Notes Comput. Sci. *8689*, 818–833.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
OCT and Chest X-Ray images and codes	https://data.mendeley.com/datasets/rscbjbr9sj/2	N/A
Software and Algorithms		
TensorFlow	https://www.tensorflow.org/	N/A
ImageNet	www.image-net.org	N/A

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and classifiers should be directed to and will be fulfilled by the Lead Contact, Kang Zhang (kang.zhang@gmail.com). There are no restrictions for use of the materials disclosed.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Images from Human Subjects

Optical coherence tomography (OCT) images (Spectralis OCT, Heidelberg Engineering, Germany) were selected from retrospective cohorts of adult patients from the Shiley Eye Institute of the University of California San Diego, the California Retinal Research Foundation, Medical Center Ophthalmology Associates, the Shanghai First People's Hospital, and Beijing Tongren Eye Center between July 1, 2013 and March 1, 2017. All OCT imaging was performed as part of patients' routine clinical care. There were no exclusion criteria based on age, gender, or race. We searched local electronic medical record databases for diagnoses of choroidal neovascularization, diabetic macular edema, drusen and normal to initially assign images. A horizontal foveal cut of OCT scans was downloaded with a standard image format according to manufacure's softwares and instructions. Chest X-ray images (anterior-posterior) were selected from retrospective cohorts of pediatric patients of one to five years old from Guangzhou Women and Children's Medical Center, Guangzhou. All chest X-ray imaging was performed as part of patients' routine clinical care. Institutional Review Board (IRB)/Ethics Committee approvals were obtained. The work was conducted in a manner compliant with the United States Health Insurance Portability and Accountability Act (HIPAA) and was adherent to the tenets of the Declaration of Helsinki.

METHOD DETAILS

OCT examinations were interpreted to confirm a diagnosis, and referral decisions were made thereafter (''urgent referral'' for diagnoses of choroidal neovascularization or diabetic macular edema, ''routine referral'' for drusen, and ''observation only'' for normal). The dataset represents the most common medical retina patients presenting and receiving treatment at all participating clinics.

Chest X-ray examinations were interpreted to confirm a diagnosis, and referral decisions were made thereafter (''urgent referral'' for diagnoses of bacterial pneumonia, ''supportive care'' for viral pneumonia, and ''observation only'' for normal).

Image Labeling

Before training, each image went through a tiered grading system consisting of multiple layers of trained graders of increasing expertise for verification and correction of image labels. Each image imported into the database started with a label matching the most recent diagnosis of the patient. The first tier of graders consisted of undergraduate and medical students who had taken and passed an OCT interpretation course review. This first tier of graders conducted initial quality control and excluded OCT images containing severe artifacts or significant image resolution reductions. The second tier of graders consisted of four ophthalmologists who independently graded each image that had passed the first tier. The presence or absence of choroidal neovascularization (active or in the form of subretinal fibrosis), macular edema, drusen, and other pathologies visible on the OCT scan were recorded. Finally, a third tier of two senior independent retinal specialists, each with over 20 years of clinical retina experience, verified the true labels for each image. The dataset selection and stratification process is displayed in a CONSORT-style diagram in Figure 2B. To account for human error in grading, a validation subset of 993 scans was graded separately by two ophthalmologist graders, with disagreement in clinical labels arbitrated by a senior retinal specialist.

For the analysis of chest X-ray images, all chest radiographs were initially screened for quality control by removing all low quality or unreadable scans. The diagnoses for the images were then graded by two expert physicians before being cleared for training the AI system. In order to account for any grading errors, the evaluation set was also checked by a third expert.

Transfer Learning Methods

Using the Tensorflow we adapted an Inception V3 architecture pretrained on the ImageNet dataset (Szegedy et al., 2016). Retraining consisted of initializing the convolutional layers with loaded pretrained weights and retraining the final, softmax layer to recognize our classes from scratch. In this study, the convolutional layers were frozen and used as fixed feature extractors. The convolutional "bottlenecks" are the values of each training and testing images after they have passed through the frozen layers of our model and since the convolutional weights are not updated, these values are initially calculated and stored in order to reduce redundant processes and speed up training. The newly initialized network, then, takes the image bottlenecks as input and retrains to classify our specific categories. Attempts at "fine-tuning" the convolutional layers by unfreezing and updating the pretrained weights on our medical images using backpropagation tended to decrease model performance due to overfitting (Figure 1).

The Inception model was trained on an Ubuntu 16.04 computer with 2 Intel Xeon CPUs, using a NVIDIA GTX 1080 8Gb GPU for training and testing, with 256Gb available in RAM memory. Training of layers was performed by stochastic gradient descent in batches of 1,000 images per step using an Adam Optimizer with a learning rate of 0.001. Training on all categories was run for 10,000 steps, or 100 epochs, since training of the final layers will have converged by then for all classes. Holdout method testing was performed after every step using a test partition containing images from patients independent of the patients represented in the training partition by passing each image through the network without performing gradient descent and backpropagation, and the best performing model was kept for analysis.

Expert Comparisons

In order to evaluate our model in the context of clinical experts, a validation set of 1000 images (633 patients), independent of the patients in the training set, was used to compare our network referral decisions with the decisions made by human experts. Weighted error scoring was used to reflect the fact that a false negative result (failing to refer) is more detrimental than a false positive result (making a referral when it was not warranted). Using these weighted penalty points, error rates were computed for the model and for each of the human experts.

Occlusion Test

Similarly to the methods described by Lee et al. and Zeiler and Fergus, an occlusion test was performed to identify the areas contributing the most to the neural network's assignment of the predicted diagnosis(Lee et al., 2016; Zeiler and Fergus, 2014). A blank 20x20 pixel box was systematically moved across every possible position in the image and the probabilities of the disease were recorded. The highest drop in the probability represents the region of interest that contributed the highest importance to the deep learning algorithm (Figure 5A, see also Figure S5 for additional examples).

QUANTIFICATION AND STATISTICAL ANALYSIS

The 207,130 images collected were reduced to the 108,312 OCT images (from 4686 patients) and used for training the AI platform. Another subset of 633 patients not in the training set was collected based on a sample size requirement of 583 patients to detect sensitivity and specificity at 0.05 marginal error and 95% confidence. The test images (n = 1000) were used to evaluate model and human expert performance. Receiver operating characteristics (ROC) curves plot the true positive rate (sensitivity) versus the false positive rate (1 – specificity). ROC curves were generated using classification probabilities of urgent referral versus otherwise and the true labels of each test image and the ROC function of the Python scikit-learn library. The area under the ROC curve is a measure of performance and the true positive rate (TPR or sensitivity) at some chosen true negative rate (TNR or specificity) on the ROC curve is the probability that the classifier will rank a randomly chosen "urgent referral" instance higher than a randomly chosen normal or drusen instance. Accuracy was measured by dividing the number of correctly labeled images by the total number of test images. Sensitivity and specificity were determined by dividing the total number of correctly labeled urgent referrals and the total number of correctly labeled non-urgent referrals, respectively, by the total number of test images.

DATA AND SOFTWARE AVAILABILITY

All deep learning methods were implemented using either TensorFlow (https://www.tensorflow.org). ImageNet, a public database of images, can be found at https://www.image-net.org. Dataset on high resolution JPEG OCT and chest X-ray images are deposited into the public Mendeley database (https://doi.org/10.17632/rscbjbr9sj.3).

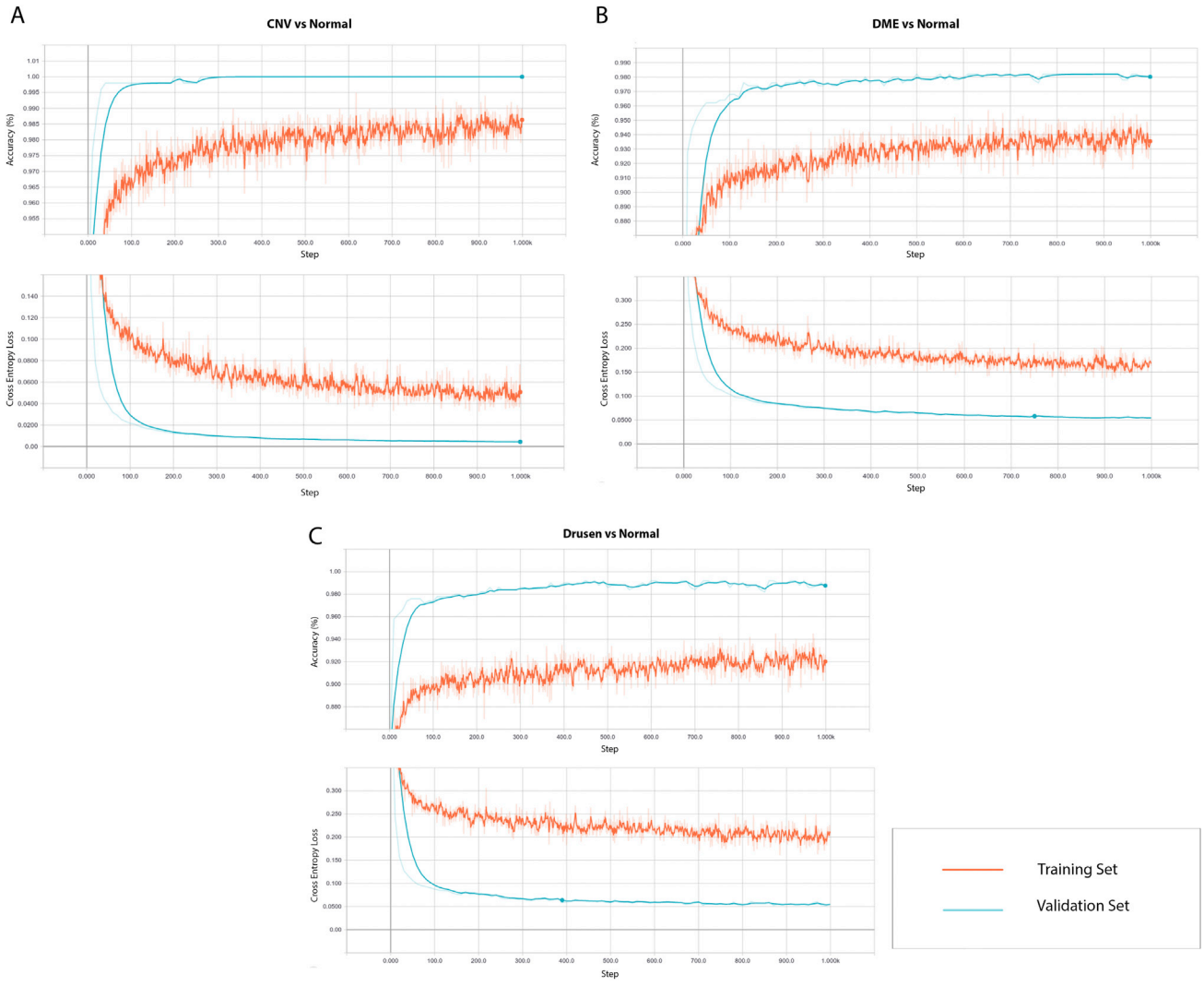


Figure S1. Plots Showing Binary Performance in the Training and Validation Datasets Using TensorBoard, Related to Figure 3

Comparisons were made for choroidal neovascularization (CNV) versus normal (A), diabetic macular edema (DME) versus normal (B), and drusen versus normal (C). Plots were normalized with a smoothing factor of 0.6 in order to clearly visualize trends. The validation accuracy and loss shows better performance since images with more noise and lower quality were also included in the training set to reduce overfitting and help generalization of the classifier. Training dataset: orange. Validation dataset: blue.

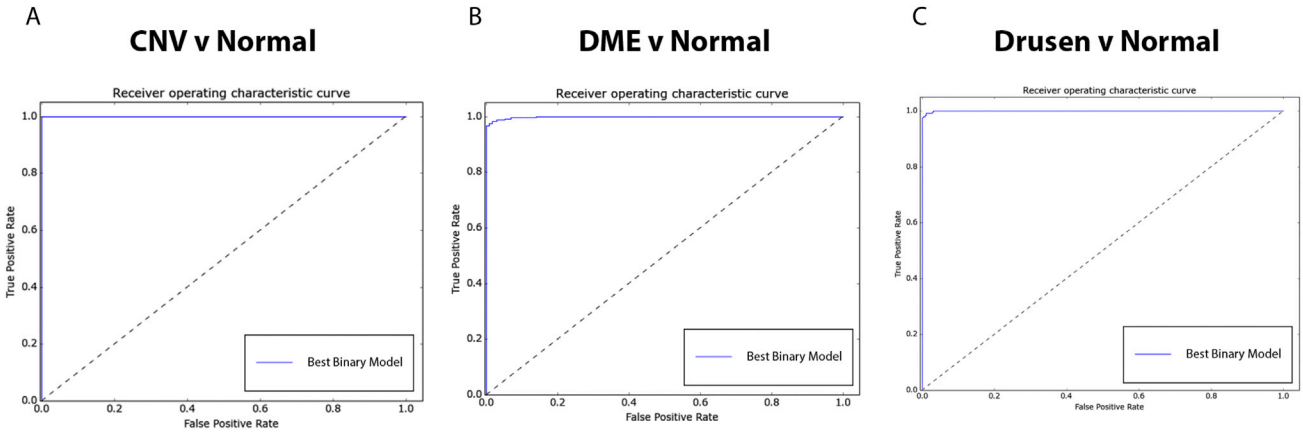


Figure S2. Receiver Operating Characteristic Curves for Binary Classifiers, Related to Figure 4

The corresponding area under the ROC curve (AUROC) for the graphs are 100% for choroidal neovascularization (CNV) versus normal (A), 99.87% for diabetic macular edema (DME) versus normal (B), and 99.96% for drusen versus normal (C). The straight vertical and horizontal lines in (A) and the nearly straight lines in (B) and (C) demonstrate that the binary convolutional neural network models have a near perfect classification performance.

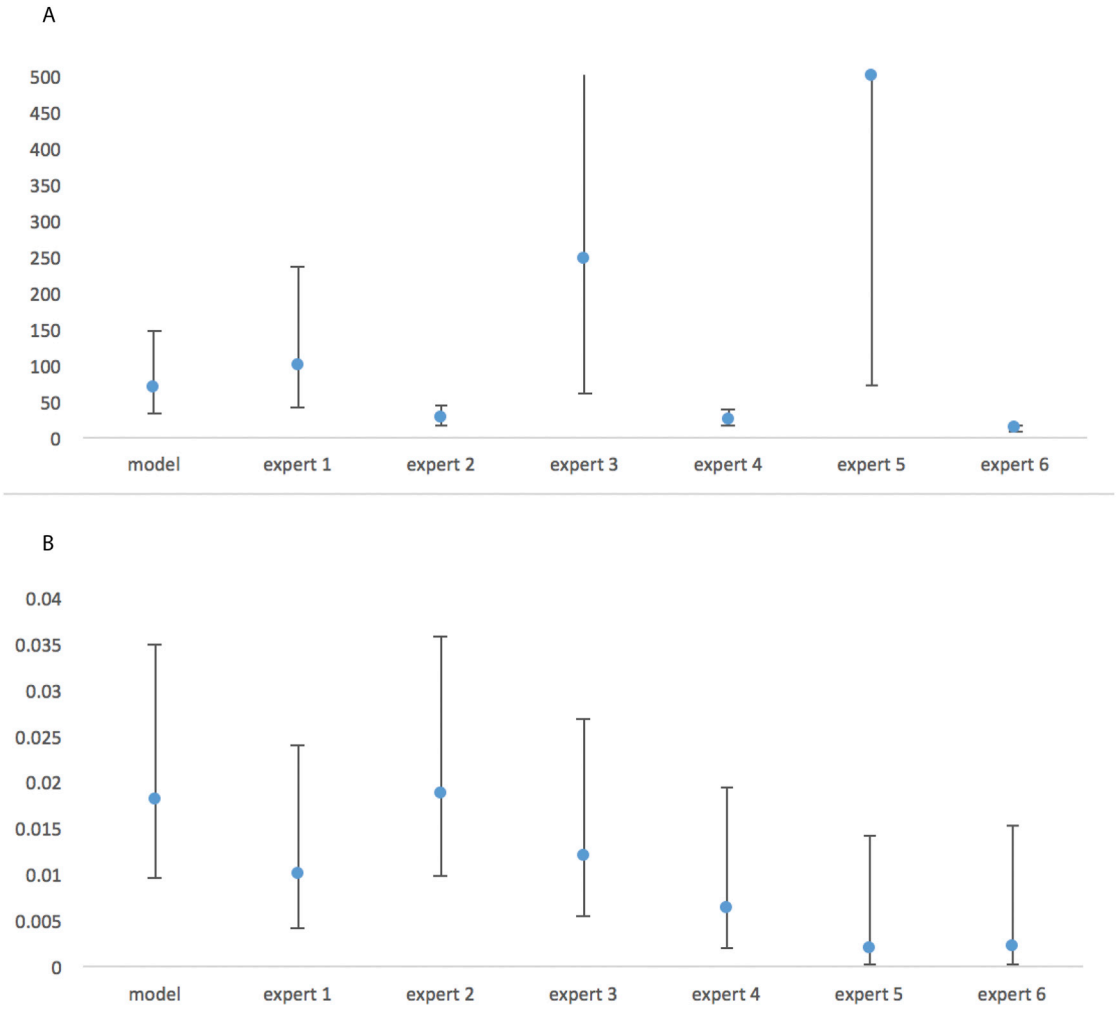


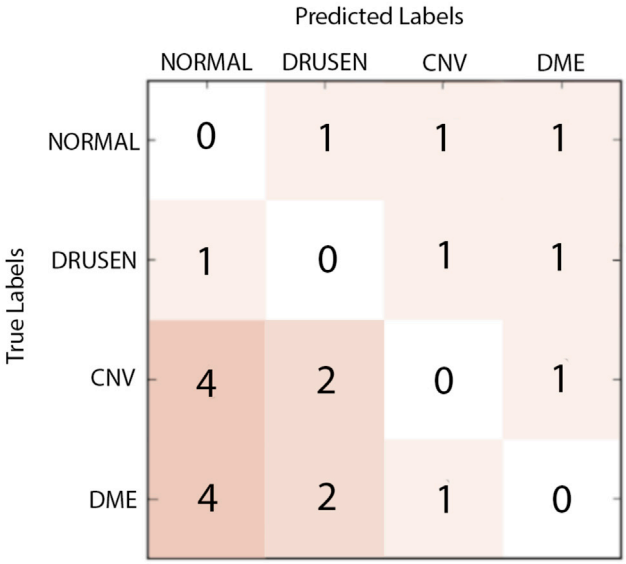
Figure S3. Plots Depicting the Positive and Negative Likelihood Ratios with Their Corresponding 95% Confidence Intervals Marked, Related to Figure 4

(A) The positive likelihood ratio is defined as the true positive rate over the false positive rate, so that an increasing likelihood ratio greater than 1 indicates increasing probability that the predicted result is associated with the disease.

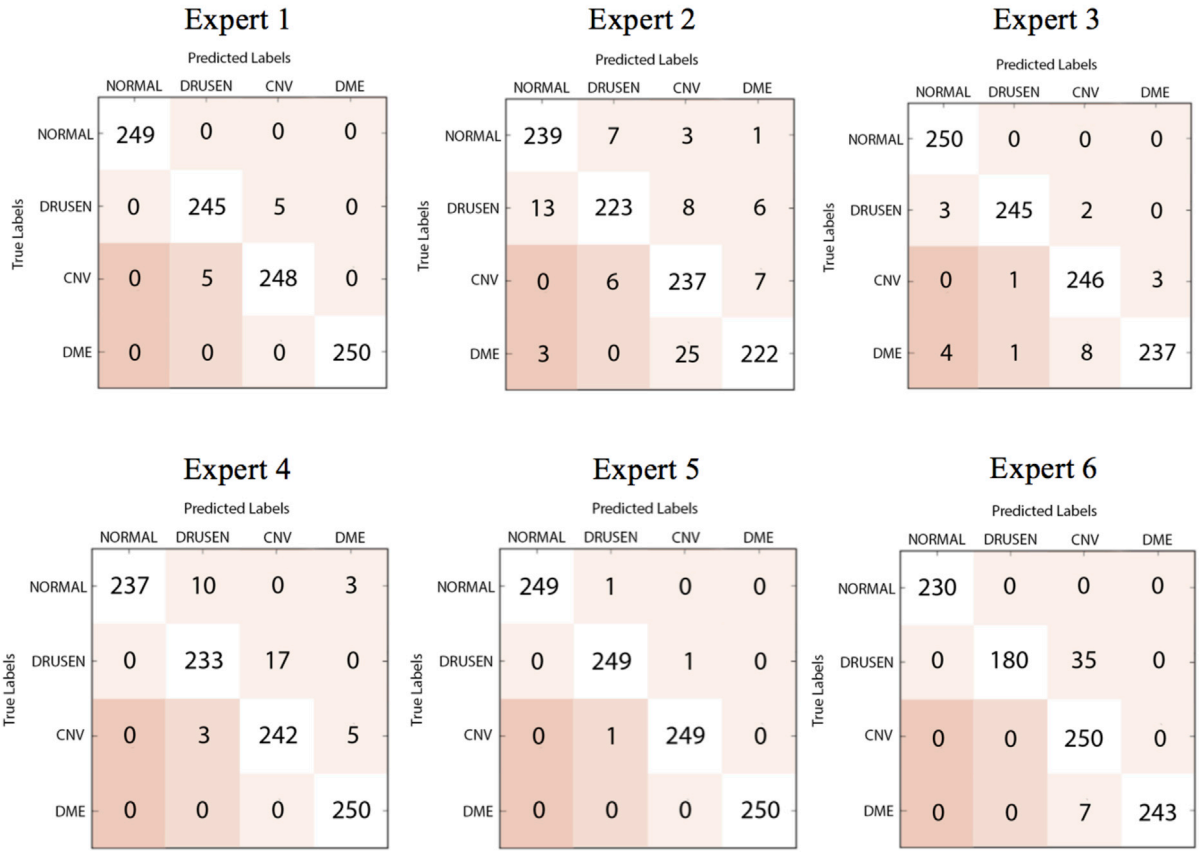
(B) The negative likelihood ratio is defined as the false negative rate over the true negative rate, so that a decreasing likelihood ratio less than 1 indicates increasing probability that the predicted result is associated with the absence of disease.

The confidence intervals show that the best trained model demonstrated statistically similar screening performance in when compared to human experts.

A



B



(legend on next page)

Figure S4. Proposed Penalties for Incorrect Labeling during Weighted Error Calculations and Confusion Matrix of Experts Grading OCT Images, Related to Figure 4

(A) The penalties include an error score of 4 for "urgent referrals" scored as normal and an error score of 2 for "urgent referrals" scored as drusen. All other incorrect answers carry an error score of 1.

(B) The results for each of the human experts is depicted here, comparing the true labels and the predicted labels for each individual grader.

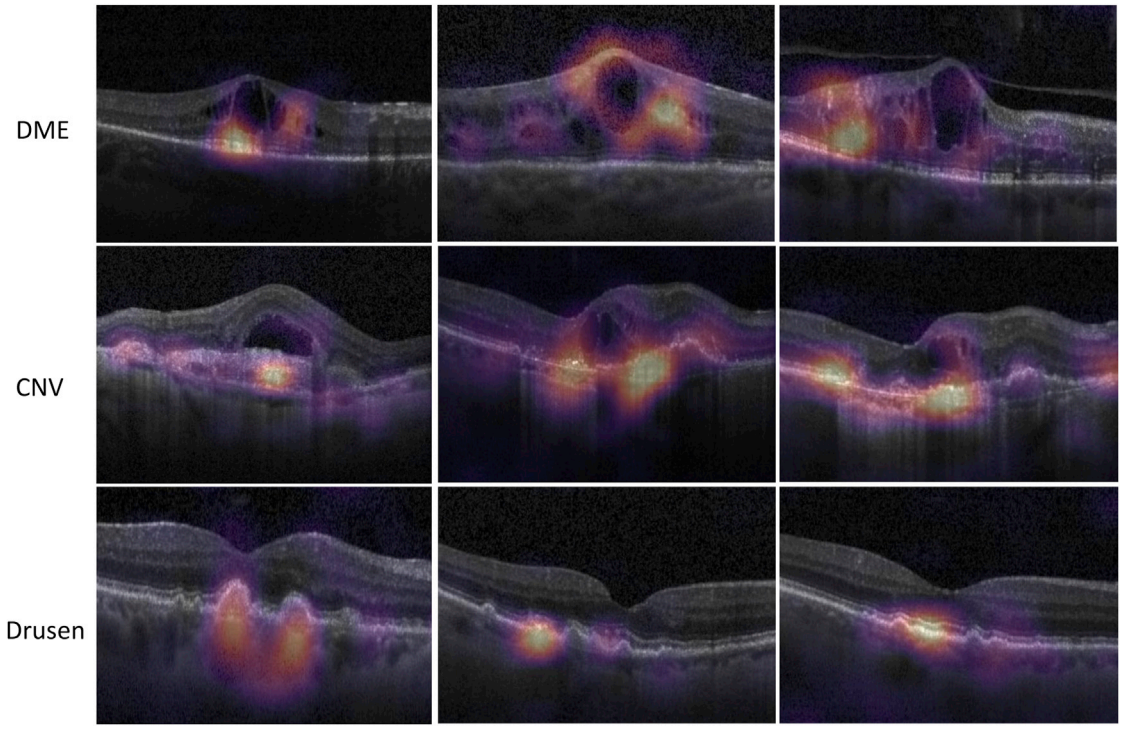


Figure S5. Occlusion Maps of Diabetic Macular Edema, Choroidal Neovascularization, and Drusen, Related to Figure 5

(Top) Diabetic macular edema (DME), (middle) choroidal neovascularization (CNV), and (bottom), drusen. Additional examples of occlusion test images, illustrating how an occluding kernel was convolved across the input image to identify areas contributing to the algorithm's determination of the diagnosis.

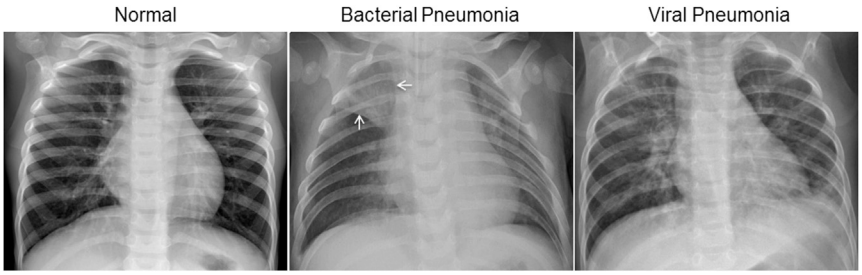


Figure S6. Illustrative Examples of Chest X-Rays in Patients with Pneumonia, Related to Figure 6

The normal chest X-ray (left panel) depicts clear lungs without any areas of abnormal opacification in the image. Bacterial pneumonia (middle) typically exhibits a focal lobar consolidation, in this case in the right upper lobe (white arrows), whereas viral pneumonia (right) manifests with a more diffuse "interstitial" pattern in both lungs.