Fully automated disease severity assessment and treatment monitoring in retinopathy of prematurity using deep learning

James M. Brown[a], J. Peter Campbell[b], Andrew Beers[a], Ken Chang[a], Kyra Donohue[a], Susan Ostmo[b], R. V. Paul Chan[c], Jennifer Dy[d], Deniz Erdogmus[d], Stratis Ioannidis[d], Michael F. Chiang[b,e], and Jayashree Kalpathy-Cramer[a,f]

[a]Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Charlestown, MA
[b]Department of Ophthalmology, Casey Eye Institute, Oregon Health & Science University, Portland, OR
[c]Department of Ophthalmology and Visualdr Sciences, Illinois Eye and Ear Infirmary, University of Illinois at Chicago, Chicago, IL
[d]Department of Electrical and Computer Engineering, Northeastern University, Boston, MA
[e]Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR
[f]Massachusetts General Hospital & Brigham and Womens Hospital Center for Clinical Data Science, Boston, MA

ABSTRACT

Retinopathy of prematurity (ROP) is a disease that affects premature infants, where abnormal growth of the retinal blood vessels can lead to blindness unless treated accordingly. Infants considered at risk of severe ROP are monitored for symptoms of plus disease, characterized by arterial tortuosity and venous dilation at the posterior pole, with a standard photographic definition. Disagreement among ROP experts in diagnosing plus disease has driven the development of computer-based methods that classify images based on hand-crafted features extracted from the vasculature. However, most of these approaches are semi-automated, which are time-consuming and subject to variability. In contrast, deep learning is a fully automated approach that has shown great promise in a wide variety of domains, including medical genetics, informatics and imaging. Convolutional neural networks (CNNs) are deep networks which learn rich representations of disease features that are highly robust to variations in acquisition and image quality. In this study, we utilized a U-Net architecture to perform vessel segmentation and then a GoogLeNet to perform disease classification. The classifier was trained on 3,000 retinal images and validated on an independent test set of patients with different observed progressions and treatments. We show that our fully automated algorithm can be used to monitor the progression of plus disease over multiple patient visits with results that are consistent with the experts' consensus diagnosis. Future work will aim to further validate the method on larger cohorts of patients to assess its applicability within the clinic as a treatment monitoring tool.

Keywords: deep learning, machine learning, retina, retinopathy of prematurity, plus disease

1. INTRODUCTION

Retinopathy of prematurity (ROP) affects premature infants and is a leading cause of childhood blindness within the United States and globally. The hallmark of the disease is abnormal growth of the retinal vessels which can lead to scarring and retinal detachment. Although most cases are mild and resolve without intervention, some infants will develop severe disease and require treatment to mitigate loss of vision. All at risk infants are monitored
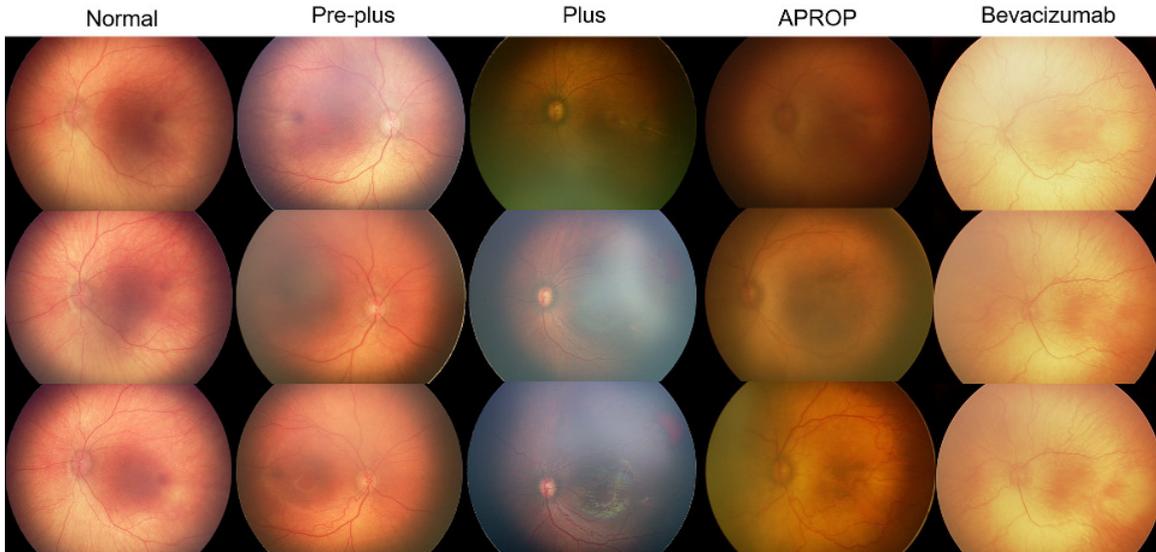
Figure 1. Example retinal photographs from each study cohort (one patient per column, consecutive visits top to bottom). Significant inter- and intra-subject variations in pathology, pigmentation, illumination, field-of-view and clarity can be observed.

weekly or biweekly, depending on disease severity, by telemedicine or clinical examination. Treatment is indicated for type-1 ROP, with aggressive posterior ROP (APROP) being a particularly severe form of the disease that rapidly progresses to the advanced stages. Treatments include laser photocoagulation therapy and intravitreal injections of anti-vascular endothelial growth factor agents (anti-VEGF) such as bevacizumab; a humanized monoclonal antibody.[1] Although there are potential advantages to anti-VEGF therapy for treatment-requiring ROP, the possibility of late disease recurrence warrants more intense monitoring of patients over the course of several months.[2]

As intensive neonatal care for premature infants continues to improve, the worldwide prevalence of ROP is predicted to increase, creating high demand for clinicians with expertise in diagnosing the disease.[3] Plus disease, characterized by increased arterial tortuosity and venous dilation,[4] is considered the most important predictor for progression of ROP.[5] Pre-plus is an intermediate level of disease that requires close monitoring for progression to plus disease.[6] High inter-rater disagreement among ROP experts has been well documented for plus disease diagnosis,[7,8] which has motivated research into automated image analysis of retinal photographs. Most methods apply computationally expensive vessel tracing algorithms to manually segmented vessel images, or manual identification of vessels and reference points, from which quantitative vessel characteristics are derived for classification.[9,10] In contrast, convolutional neural networks (CNNs) are fully automated and do not require hand-crafted features. Recent work has shown the robustness of CNNs to a wide range of medical imaging problems.[11,12] CNNs have been applied previously to retinal diseases including diabetic retinopathy,[13] macular degeneration[14] and binary classification of plus disease.[15] In this study, we present the results of a fully automated algorithm based on CNNs for automated diagnosis of plus disease across multiple patient visits, demonstrating its potential as a monitoring tool for ROP.

2. DATA

This study was approved by the Institutional Review Board at the coordinating center (Oregon Health & Science University [OHSU]) and at each of 7 study centers (Columbia University, University of Illinois at Chicago, William Beaumont Hospital, Childrens Hospital Los Angeles, Cedars-Sinai Medical Center, University of Miami, Asociacion para Evitar la Ceguera en Mexico [APEC]), and followed the tenets of the Declaration of Helsinki. Written informed consent was obtained from parents of all infants enrolled.
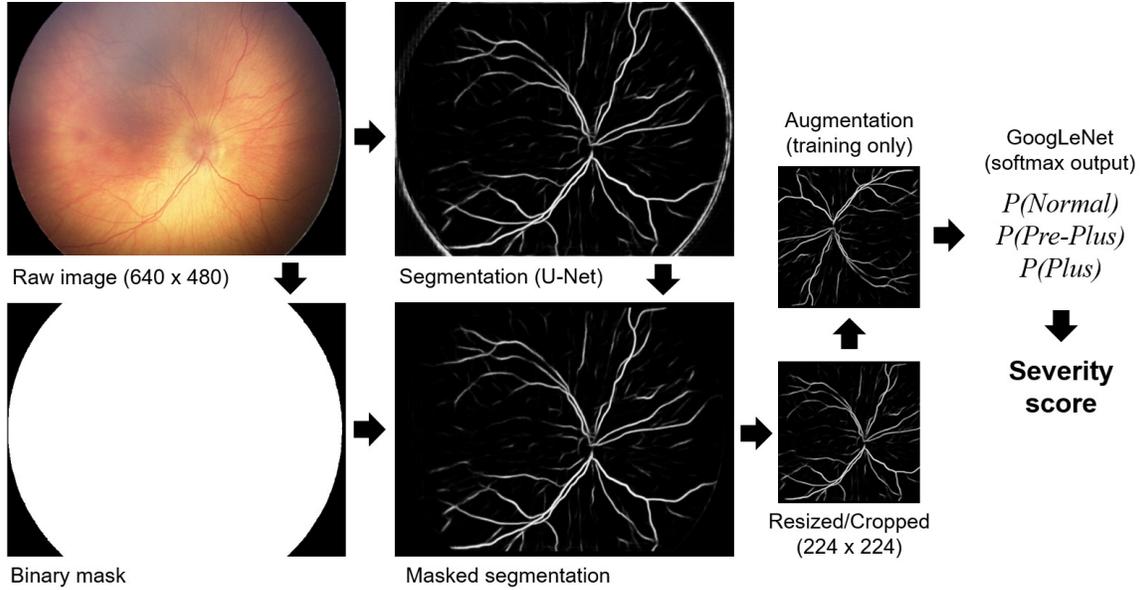
Figure 2. Diagram of the pre-processing and classification pipeline to estimate plus disease severity.

A training dataset was created from a database of 4,800 de-identified posterior retinal images captured using a commercially available camera (RetCam; Natus Medical Incorporated, Pleasanton, CA). These data were collected as part of the multicenter Imaging and Informatics in ROP (i-ROP) cohort study. A reference standard diagnosis (RSD) was assigned to each image in the training set using previously published methods,[16] which served as the ground truth for CNN training. For testing, a dataset of 131 images from 46 patients were gathered and divided into five cohorts based on the RSDs observed over multiple visits; *no progression*, *progression to pre-plus*, *progression to plus*, *progression to APROP*, and *bevacizumab-treated*. No images from patients in the testing set were included in the training set. Representative example images from each cohort are shown in Figure 1.

3. METHODS

3.1 Pre-processing

Figure 2 shows a diagram of the analysis pipeline. For preprocessing, all images were resized to 640 x 480 pixels. A patch-based (48 x 48 pixel) U-Net[17] CNN architecture was then trained on 200 manually segmented retinal images, minimizing a cross-entropy loss function with stochastic gradient descent and a learning rate of 0.01. The trained U-Net was then used to segment the entire training set by extracting all overlapping patches with an 8 pixel stride. Overlapping regions were averaged to produce an image in the range 0.0 - 1.0. An approximate binary mask was computed by applying a simple threshold to the original image. This approximation was then used to calculate the centroid and radius of a circle enclosing the retinal field of view. This refined mask was applied to the segmentation image to remove incorrectly segmented regions around the periphery. The training images were then cropped and resized to 224 x 224 pixels with bicubic interpolation. To balance the training set, the images were augmented by means of 90 degree rotations and horizontal/vertical flips, increasing the total number of images eight-fold. A subset of 3,024 images (1,008 from each class) was randomly sampled from the augmented set and used for CNN training.

3.2 Severity score

The Inception-v1 (GoogLeNet)[18] CNN architecture was used for image classification. The weights were initialized following training on the ImageNet[19] database (~1.2 million images); a strategy commonly referred to as *transfer learning*.[20] The final softmax activation was modified to produce three outputs; *normal*, *pre-plus* or *plus*. The

network was then trained for 100 epochs with a constant learning rate of 0.0001 using stochastic gradient descent, with cross-entropy loss. These hyperparameters were tuned by cross-validation within the training set. Inference was performed on the testing images following pre-processing, yielding probabilities for each of the three disease classes. A severity score was then computed as a weighted sum of the probabilities

$$S = P(Normal) + 2P(PrePlus) + 3P(Plus) \tag{1}$$

ranging from 1.0 (normal) to 3.0 (plus). This severity score was used to assess the same eye patients over multiple visits. The elapsed time between patient visits was not identical.

3.3 Feature visualization

To visualize changes in appearance due to pathology over multiple patient visits, t-distributed stochastic neighbor embedding (t-SNE) was used to map the high-dimensional CNN features into a two-dimensional latent space.[21] The fully-connected network at the end of the GoogLeNet architecture was removed, and the output of the global average pooling layer used to extract a vector of image features $\mathbf{x} \in \mathbb{R}^{1024}$ for each image in the test set. t-SNE was then applied to the features of the test set $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}\}, \mathbf{x_i} \in \mathbb{R}^{1024}$ to produce an embedding $\mathbf{Y} = \{\mathbf{y_1}, \mathbf{y_2}, ..., \mathbf{y_n}\}, \mathbf{y_i} \in \mathbb{R}^2$, which was visualized as a scatter plot.

4. RESULTS AND DISCUSSION

4.1 Severity scores over time

Figure 3 shows the results of automated diagnosis over multiple visits for the five patient cohorts. Plots show the mean and standard deviation severity score at each visit. Infants reported as having no ROP progression (n = 6, 3 or 4 visits) show no noticeable signs of disease across all visits, with negligible variance across patients. The Pre-Plus cohort shows a steady increase to intermediate severity, whereas the Plus (n = 8, 3 or 4 visits) and APROP (n = 6, 2 or 3 visits) patients both exhibit a more pronounced increase to maximal severity. The rate of progression in APROP appears to be both rapid and more varied between patients across all visits. Bevacizumab-treated infants (n = 15, 2 or 3 visits) show regression after treatment (between visits 0 and 1) followed by varying degrees of progression at visit 2. This is consistent with trends observed clinically following bevacizumab treatment.[2, 22]
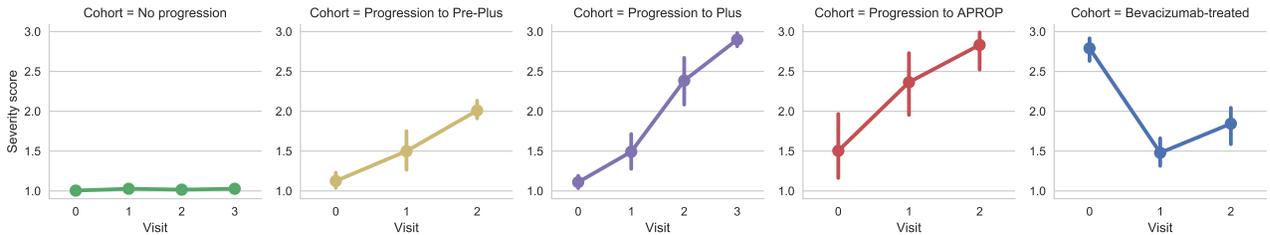

Figure 3. Mean disease severity plotted for five patient cohorts over multiple visits. The severity score is calculated based on the output probabilities of the CNN (ranging from 1.0 3.0). Error bars correspond to one standard deviation.

4.2 Disease trajectories in feature space

The t-SNE disease trajectories are shown in Figure 4, colorized by patient cohort. The bevacizumab-treated cohort is omitted for clarity. The size of individual points in the space is proportional to the severity score for that image. Images from the same patient are connected by vectors that denote disease trajectories over time. It is apparent that the t-SNE latent space separates disease stages, with non-pathological images clustered in the upper left and severe pathology on the lower right.
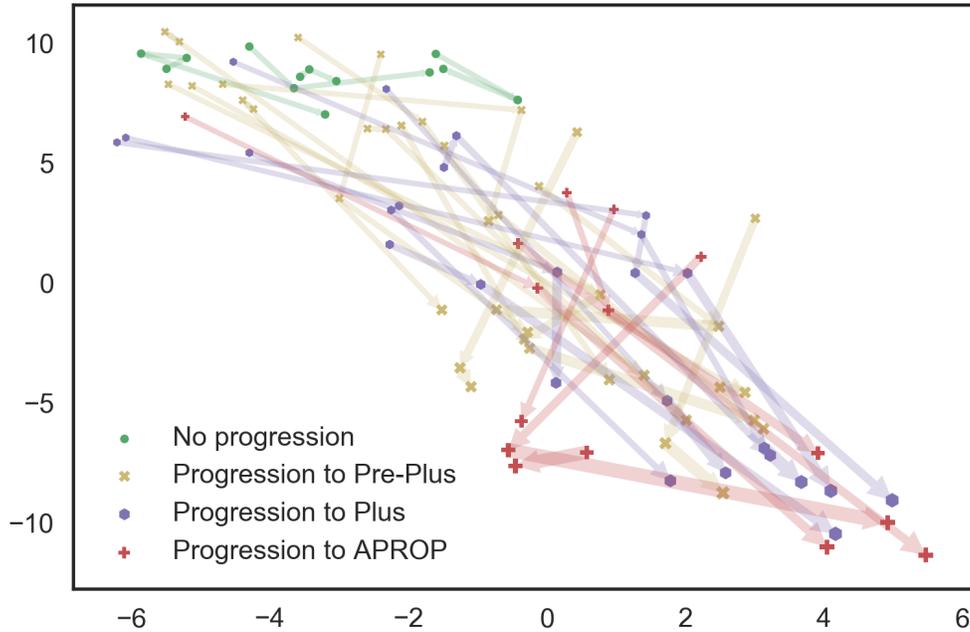
Figure 4. Disease trajectory plots in t-SNE latent space (arbitrary units). Each point corresponds to in an individual image, with individual patients indicated by a set of connected vectors. Vector thickness is proportional to disease severity. The bevacizumab-treated cohort is omitted for clarity.

5. CONCLUSIONS

In this study, we present a deep learning-based algorithm for monitoring plus disease in ROP. Application of the method to a small dataset of patients with different disease characteristics yields promising results, showing trends consistent with the reference standard diagnoses. t-SNE visualization of learned image features allows for direct comparison of images and patients in a common space. To our knowledge, this is the first application of a fully automated method for monitoring ROP over time. Future work will focus on application of the method to larger cohorts of patients, and interpretation of the features learned by the algorithm. The proposed method shows potential for earlier identification of patients that are likely to progress to ROP and assist clinical decision-making.

ACKNOWLEDGMENTS

REFERENCES

[1] Mintz-Hittner, H. A., Kennedy, K. A., Chuang, A. Z., and BEAT-ROP Cooperative Group, "Efficacy of Intravitreal Bevacizumab for Stage 3+ Retinopathy of Prematurity," *New England Journal of Medicine* **364**(7), 603–615 (2011).

[2] Mintz-Hittner, H. A., Geloneck, M. M., and Chuang, A. Z., "Clinical Management of Recurrent Retinopathy of Prematurity after Intravitreal Bevacizumab Monotherapy," *Ophthalmology* **123**(9), 1845–1855 (2016).

[3] Gilbert, C., Rahi, J., Eckstein, M., O'Sullivan, J., and Foster, A., "Retinopathy of prematurity in middle-income countries," *The Lancet* **350**(9070), 12–14 (1997).

[4] Early Treatment for Retinopathy of Prematurity Cooperative Group, "Revised Indications for the Treatment of Retinopathy of Prematurity," *Archives of Ophthalmology* **121**(12), 1684 (2003).

[5] Cryotherapy for Retinopathy of Prematurity Cooperative Group, "Multicenter trial of cryotherapy for retinopathy of prematurity: preliminary results," *Arch Ophthalmol* **106**(4), 471–479 (1988).

[6] Good, W. V., "Final results of the Early Treatment for Retinopathy of Prematurity (ETROP) randomized trial.," *Transactions of the American Ophthalmological Society* **102**, 233–48; discussion 248–50 (2004).

[7] Chiang, M. F., Jiang, L., Gelman, R., Du, Y. E., and Flynn, J. T., "Interexpert agreement of plus disease diagnosis in retinopathy of prematurity.," *Archives of ophthalmology* **125**(7), 875–80 (2007).

[8] Kalpathy-Cramer, J., Campbell, J. P., Erdogmus, D., Tian, P., Kedarisetti, D., Moleta, C., Reynolds, J. D., Hutcheson, K., Shapiro, M. J., Repka, M. X., Ferrone, P., Drenser, K., Horowitz, J., Sonmez, K., Swan, R., Ostmo, S., Jonas, K. E., Chan, R. V. P., Chiang, M. F., Chiang, M. F., Ostmo, S., Sonmez, K., Campbell, J. P., Chan, R. V. P., Jonas, K., Horowitz, J., Coki, O., Eccles, C. A., Sarna, L., Berrocal, A., Negron, C., Denser, K., Cumming, K., Osentoski, T., Check, T., Zajechowski, M., Lee, T., Kruger, E., McGovern, K., Simmons, C., Murthy, R., Galvis, S., Rotter, J., Chen, I., Li, X., Taylor, K., Roll, K., Kalpathy-Cramer, J., Erdogmus, D., Martinez-Castellanos, M. A., Salinas-Longoria, S., Romero, R., Arriola, A., Olguin-Manriquez, F., Meraz-Gutierrez, M., Dulanto-Reinoso, C. M., and Montero-Mendoza, C., "Plus Disease in Retinopathy of Prematurity: Improving Diagnosis by Ranking Disease Severity and Using Quantitative Image Analysis," *Ophthalmology* **123**(11), 2345–2351 (2016).

[9] Wallace, D. K., "Computer-assisted quantification of vascular tortuosity in retinopathy of prematurity (an American Ophthalmological Society thesis).," *Transactions of the American Ophthalmological Society* **105**, 594–615 (2007).

[10] Ataer-Cansizoglu, E., Bolon-Canedo, V., Campbell, J. P., Bozkurt, A., Erdogmus, D., Kalpathy-Cramer, J., Patel, S., Jonas, K., Chan, R. V. P., Ostmo, S., and Chiang, M. F., "Computer-Based Image Analysis for Plus Disease Diagnosis in Retinopathy of Prematurity: Performance of the i-ROP System and Image Features Associated With Expert Diagnosis," *Translational Vision Science & Technology* **4**(6), 5 (2015).

[11] LeCun, Y., Bengio, Y., and Hinton, G., "Deep learning.," *Nature* **521**(7553), 436–44 (2015).

[12] Ravi, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., and Yang, G. Z., "Deep Learning for Health Informatics," *IEEE Journal of Biomedical and Health Informatics* **21**(1), 4–21 (2017).

[13] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., and Webster, D. R., "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," *JAMA* **316**(22), 2402 (2016).

[14] Lee, C. S., Baughman, D. M., and Lee, A. Y., "Deep Learning Is Effective for Classifying Normal versus Age-Related Macular Degeneration OCT Images," *Ophthalmology Retina* **1**(4), 322–327 (2017).

[15] Worrall, D. E., Wilson, C. M., and Brostow, G. J., "Automated Retinopathy of Prematurity Case Detection with Convolutional Neural Networks," in [*International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*], (2), 68–76 (2016).

[16] Ryan, M. C., Ostmo, S., Jonas, K., Berrocal, A., Drenser, K., Horowitz, J., Lee, T. C., Simmons, C., Martinez-Castellanos, M. A., Chan, R. V. P., and Chiang, M. F., "Development and Evaluation of Reference Standards for Image-based Telemedicine Diagnosis and Clinical Research Studies in Ophthalmology," in [*AMIA Annu Symp Proc.*], **2014**, 1902–1910, American Medical Informatics Association (2014).

[17] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional networks for biomedical image segmentation," *Lecture Notes in Computer Science* **9351**, 234–241 (2015).

[18] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A., "Going deeper with convolutions," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **07-12-June**, 1–9 (2015).

[19] Jia Deng, Wei Dong, Socher, R., Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in [*2009 IEEE Conference on Computer Vision and Pattern Recognition*], 248–255, IEEE (2009).

[20] Yosinski, J., Clune, J., Bengio, Y., and Lipson, H., "How transferable are features in deep neural networks?," *Advances in Neural Information Processing Systems 27 (Proceedings of NIPS)* **27**, 1–9 (2014).

[21] Maaten, L. V. D. and Hinton, G., "Visualizing Data using t-SNE," *Journal of Machine Learning Research* **620**(1), 267–84 (2008).

[22] Hu, J., Blair, M. P., Shapiro, M. J., Lichtenstein, S. J., Galasso, J. M., and Kapur, R., "Reactivation of retinopathy of prematurity after bevacizumab injection," *Archives of Ophthalmology* **130**(8), 1000–1006 (2012).