tvst

Special Issue

Plus Disease in Retinopathy of Prematurity: Convolutional Neural Network Performance Using a Combined Neural Network and Feature Extraction Approach

Veysi M. Yildiz[1,*], Peng Tian[1], Ilkay Yildiz[1], James M. Brown[2], Jayashree Kalpathy-Cramer[3], Jennifer Dy[1], Stratis Ioannidis[1], Deniz Erdogmus[1], Susan Ostmo[4], Sang Jin Kim[5], R. V. Paul Chan[6], J. Peter Campbell[4], and Michael F. Chiang[4], for the Imaging and Informatics in Retinopathy of Prematurity (i-ROP) Research Consortium

[1] Cognitive Systems Laboratory, Northeastern University, Boston, MA, USA
[2] Department of Computer Science, University of Lincoln, Lincoln, UK
[3] Department of Radiology, Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, MA, USA
[4] Department of Ophthalmology, Casey Eye Institute, Oregon Health & Science University, Portland, OR, USA
[5] Sungkyunkwan University School of Medicine, Seoul, South Korea
[6] Eye and Ear Infirmary, University of Illinois at Chicago, Chicago, IL, USA

Correspondence: Veysi M. Yildiz, Cognitive Systems Laboratory, Northeastern University, 360 Huntington Ave, 409 Dana, Boston, MA 02115, USA. e-mail: yildiz@ece.neu.edu

Received: April 26, 2019
Accepted: August 5, 2019
Published: February 14, 2020

Keywords: ROP; CNN; feature-based

Citation: Yildiz VM, Tian P, Yildiz I, Brown JM, Kalpathy-Cramer J, Dy J, Ioannidis S, Erdogmus D, Ostmo S, Kim SJ, Chan RVP, Campbell JP, Chiang MF, for the Imaging and Informatics in Retinopathy of Prematurity (i-ROP) Research Consortium. Plus disease in retinopathy of prematurity: convolutional neural network performance using a combined neural network and feature extraction approach. Trans Vis Sci Tech. 2020;9(2):10, https://doi.org/10.1167/tvst.9.2.10

Purpose: Retinopathy of prematurity (ROP), a leading cause of childhood blindness, is diagnosed by clinical ophthalmoscopic examinations or reading retinal images. Plus disease, defined as abnormal tortuosity and dilation of the posterior retinal blood vessels, is the most important feature to determine treatment-requiring ROP. We aimed to create a complete, publicly available and feature-extraction-based pipeline, I-ROP ASSIST, that achieves convolutional neural network (CNN)-like performance when diagnosing plus disease from retinal images.

Methods: We developed two datasets containing 100 and 5512 posterior retinal images, respectively. After segmenting retinal vessels, we detected the vessel centerlines. Then, we extracted features relevant to ROP, including tortuosity and dilation measures, and used these features in the classifiers including logistic regression, support vector machine and neural networks to assess a severity score for the input. We tested our system with fivefold cross-validation and calculated the area under the curve (AUC) metric for each classifier and dataset.

Results: For predicting plus versus not-plus categories, we achieved 99% and 94% AUC on the first and second datasets, respectively. For predicting pre-plus or worse versus normal categories, we achieved 99% and 88% AUC on the first and second datasets, respectively. The CNN method achieved 98% and 94% for predicting two categories on the second dataset.

Conclusions: Our system combining automatic retinal vessel segmentation, tracing, feature extraction and classification is able to diagnose plus disease in ROP with CNN-like performance.

Translational Relevance: The high performance of I-ROP ASSIST suggests potential applications in automated and objective diagnosis of plus disease.

Introduction

Retinopathy of prematurity (ROP) is a disease that can be diagnosed from findings either by clinical ophthalmoscopic examinations or reading fundus images. It mostly affects infants with a birth weight of ≤1500 g or gestational age of 30 weeks or less.[1] ROP is characterized by aberrant retinal vascular development, vascular abnormalities, and neovascularization. In mild ROP, retinal vascular pathologies regress spontaneously, but in severe cases the vascular abnormalities progress to fibrovascular proliferation and retinal detachment with subsequent blindness. ROP remains one of the leading causes of childhood blindness in the world.[2] Furthermore, as the survival rate of premature infants is increasing, the number of infants at the risk of ROP is increasing.[3]

To standardize ROP diagnosis, an international classification system was developed in the 1980s and revised in 2005.[4,5] According to this system, plus disease is the most important disease feature in determining the need for treatment, and it is defined as abnormal tortuosity and dilation of the posterior retinal blood vessels. The system developed in the 1980s defined plus disease as a binary variable (i.e., present or absent). However, in 2005, a refined system defined a new class, pre-plus. According to this definition, a pre-plus eye has vascular abnormalities in the retina, yet dilation and tortuosity of the retinal vessels are insufficient to label the corresponding eye with plus disease.[5] The presence of plus disease indicates that treatment such as laser photocoagulation is appropriate.[1] Therefore, it is important to diagnose plus disease accurately.

There have been multiple attempts to automate the measurement of the vascular dilation and tortuosity that constitute plus disease, to variable degrees of success. Most of the current retinal image analysis systems are not fully automated. CAIAR,[6] RISA[7] and the system proposed by Fiorin and Ruggeri[9] are semi-automatic systems for quantifying dilation and tortuosity of retinal vessels. Also, RIVERS[8] traces retinal vessel centerlines and detects the changes between registered images.

Recently, there have been several descriptions of automated systems. There are two categories of automated systems: (1) systems that extract handcrafted features (e.g., vascular tortuosity and diameter of the vessels), and (2) systems that employ convolution neural networks (CNNs) for feature extraction. The algorithm developed by Pour et al.[10] and ROPTool[11] use handcrafted features for ROP detection. ROPTool determines the existence of ROP by comparing the values of handcrafted features in each quadrant with a predefined average. On the other hand, Brown et al.[12,13] have employed CNN for segmenting retina vessels, diagnosing plus disease and monitoring ROP treatment. Their CNN approach achieves 0.94 and 0.98 area under the curve (AUC) statistics for predicting normal versus pre-plus or worse and plus versus not-plus disease, respectively. Also, Worrall et al.[14] proposed a CNN method for detecting plus disease that achieved 92% accuracy. In many medical imaging tasks, CNNs have been found to have improved performance compared with feature-extraction-based machine learning approaches[15]; however, they have the limitation that the CNN features are not transparent or explainable.

In this paper, we combine some of the advantages of a CNN model for identification of the relevant vascular structures with a feature-extraction algorithm previously developed[16] to determine whether combining these models might produce an automated plus disease classifier with performance similar to that of CNNs but with explainable features. The I-ROP ASSIST system is inspired by that of Ataer-Cansizoglu,[16] who proposed a system to create a severity score for a retina image. Even though we follow that system's vessel tracing and feature extraction methods, our system differs in several aspects. We add an optic disc center detector to make the system fully automated and replace the vessel segmentation method. We compare three different classifiers to show that our handcrafted features are discriminative for detection of plus disease. Finally, our pipeline is a freely available package written in Python and does not require a paid license to use. The package is accessible at https://github.com/neu-spiral/iROPASSISTPackage.

Methods

Dataset

This study was approved by the Institutional Review Board at the coordinating center (Oregon Health & Science University) and at each of eight study centers (Columbia University, University of Illinois at Chicago, William Beaumont Hospital, Children's Hospital Los Angeles, Cedars-Sinai Medical Center, University of Miami, Weill Cornell Medical Center and Asociacion para Evitar la Ceguera en Mexico). This study was conducted in accordance with the Declaration of Helsinki. Written informed consent for the study was obtained from parents of all infants enrolled.

As part of the Imaging and Informatics in ROP (i-ROP) study, a multicenter ROP cohort study, we

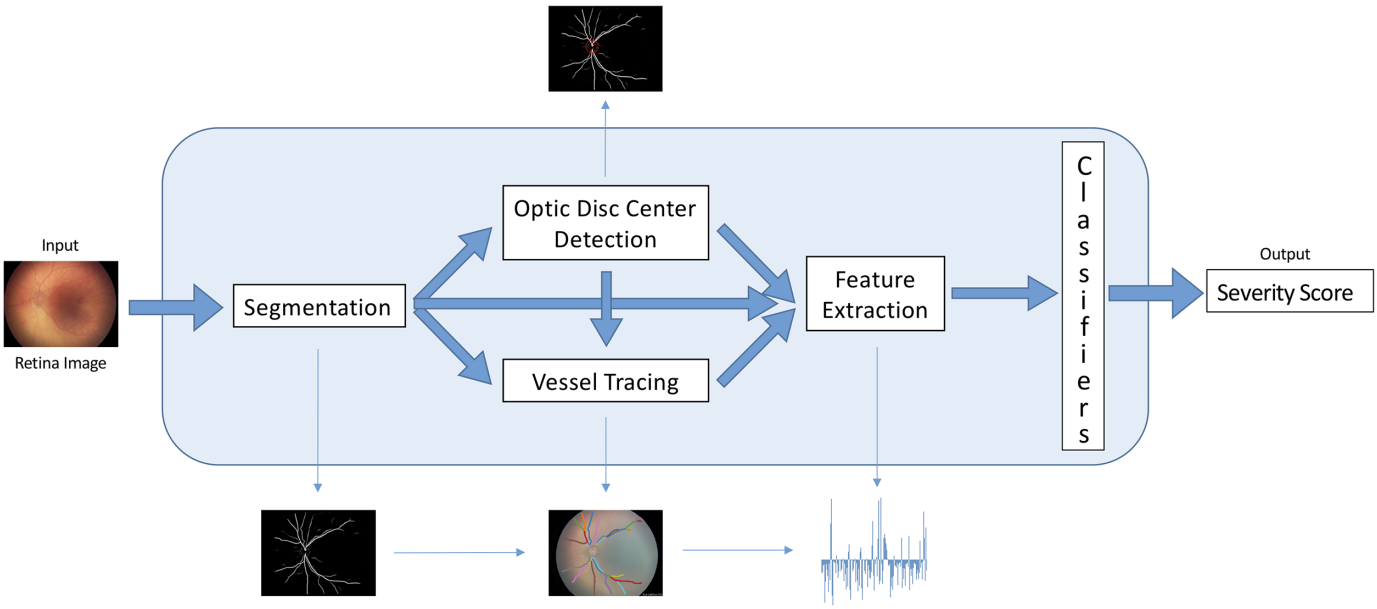

Figure 1. System pipeline. Input retina image first goes through vessel segmentation process, and the optic disc center is then detected. Using segmented images and optic disc centers, the vessels are traced and vessel tree information is extracted. Using the outputs from previous steps, features of the retina are extracted, and these features are used for classification.

developed two datasets containing 100 and 5512 posterior retinal images, respectively. The retinal images were taken using a RetCam® wide-angle fundus camera (Natus Medical Inc., Pleasanton, CA, USA) between July 2011 and December 2016. A reference standard diagnosis (i.e., plus, pre-plus or normal) was assigned to each of the images as previously described.[17] In brief, the reference standard diagnosis was established based on the consensus diagnosis that combined the image-based diagnosis by three independent expert graders and the clinical diagnosis at each study center. The large dataset containing 5512 images consisted of 163 plus, 802 pre-plus, and 4547 normal images and was used for training and validation using a cross-validation approach. The 100-image dataset, which was used by Ataer-Cansizoglu,[16] contained 15 plus, 34 pre-plus, and 51 normal images. The smaller dataset was a fully independent dataset that has been well characterized by multiple expert classifications in prior work.[18–20] It is included in this study to present the performance of our system on previously studied dataset.

System Pipeline

We divide the plus disease diagnosis procedure into 5 steps, as shown in Figure 1. First, we take color retinal images as inputs, segment the vessels, and find the optic disc center, the center point of optic disc where the optic nerve fibers leave the retina. Second, we detect centerlines of the vessels and the vessel tree structure and then extract 143 features based on dilation and tortuosity of the vessels. Finally, we produce the severity score via a classifier. Segmented single-channel images and optic disc centers are optional inputs to the system. If they are provided, the system runs the remaining steps based on the provided inputs. These steps are independent and can be modified for future improvements.

Segmentation

The pipeline of the system begins with segmenting the vessels in a color retina image. We followed the segmentation procedure from Brown et al.[12] The system deploys a pre-trained U-Net CNN[21] architecture for segmentation. The patch size of the architecture is 48 × 48. The U-Net CNN was trained on 200 manually segmented images, and cross-entropy loss function was employed. Here, stochastic gradient descent with a learning rate of 0.01 was used as the minimizer. The trained network segments the input image by extracting all overlapping patches with an 8-pixel stride. Using 8-pixel strides results in overlapping regions, and these regions are averaged to produce an output image with pixel values ranging from 0 to 1. The resulting image contains a circle enclosing the retinal field of view. We removed the circle with a mask obtained by applying a threshold to the image and estimating the center and radius of the circle. The final output is a 640 × 480, single-channel (gray) image of

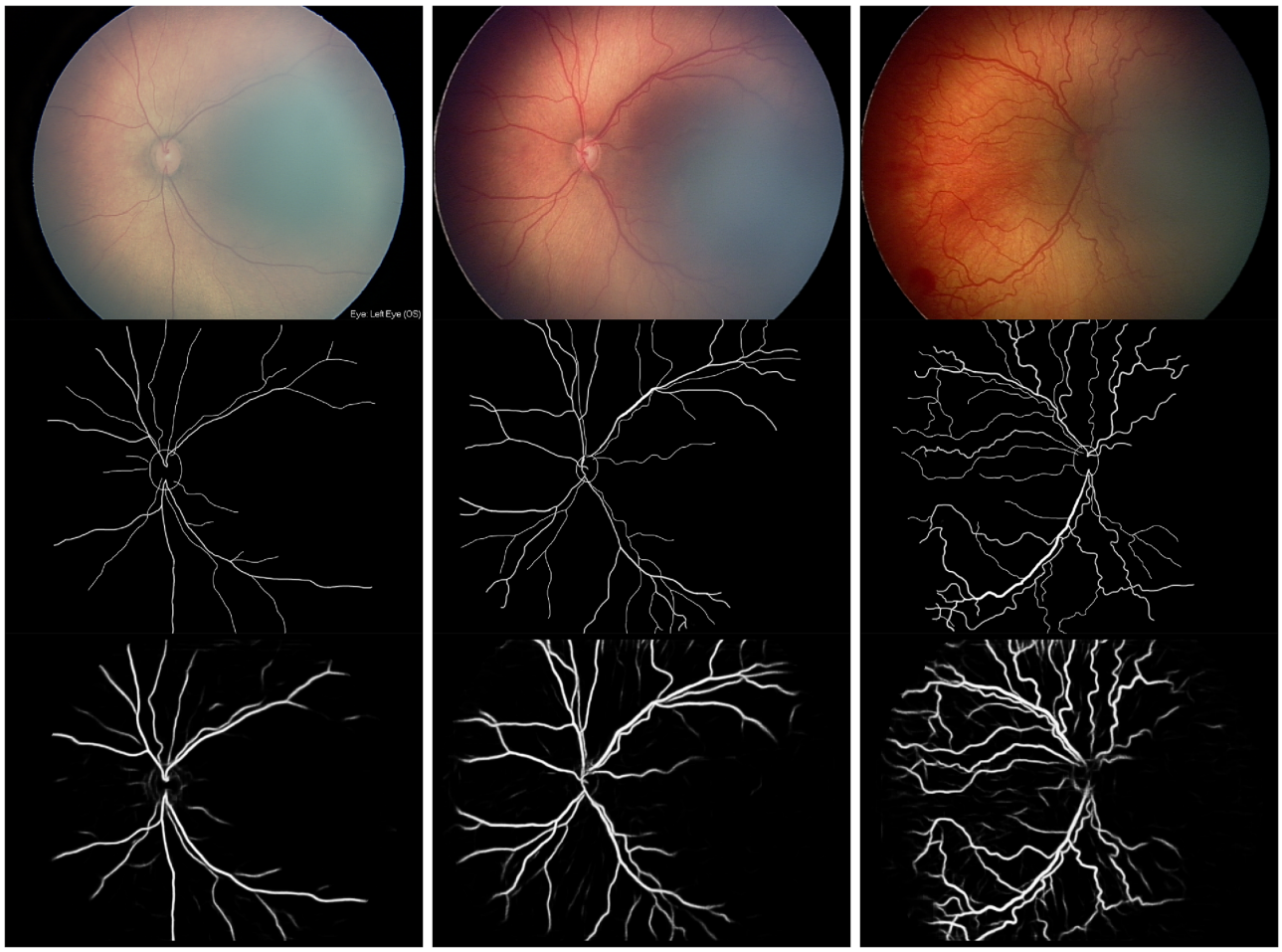

Figure 2. Images in the first row are the input color retina images, in the middle row are corresponding manual segmentations, and in the last row are the resultant automatically segmented images. Also, beginning from the left column, images are ordered according to their severity level (i.e., normal, pre-plus, and plus).

the vessels in the retina. The output of the U-Net CNN can be interpreted as an image, pixel values of which are the probability of being a vessel.

Figure 2 shows sample retina images and their segmented versions. Images in the first row are the input retina images, in the middle row are the corresponding manual segmentations used by Ataer-Cansizoglu et al.,[16,19] and in the last row are the resultant automatically segmented images using the Brown et al.[12] method. Also, beginning from the left column, images are ordered according to their severity level (i.e., normal, pre-plus, and plus).

Optic Disc Center Detection

In our system, the optic disc center is used in both vessel tracing and feature extraction modules. When creating the vessel tree structure of the image, the system begins with the vessels that are connected to the optic disc (i.e., a circle around the optic disc center with a radius of 30 pixels). Also, in the feature extraction module there are several features that use the optic disc center for distance measures. In our system, the user can provide the optic disc center as an input, or the system can detect the optic disc center automatically. We show the performances of both cases in the Results section. The automatic optic disc center detector employs a CNN that consists of the downsampling arm of U-Net.[21] The CNN was trained on 5000 segmented images. The loss function deployed for training was the Euclidean distance between the predicted disc center and ground truth, provided by ROP experts. The loss function can be formulated as

$$L(\boldsymbol{c}, \hat{\boldsymbol{c}}) = \|\boldsymbol{c} - \hat{\boldsymbol{c}}\|_2 \quad (1)$$

where $\boldsymbol{c}$ is the true disc center and $\hat{\boldsymbol{c}}$ is the predicted disc center.

Table 1. Segment-Based Features and Their Corresponding Formulations

Feature	Formula
Cumulative tortuosity index (*CTI*)	$CTI(v) = L_c(v)/L_x(v)$
Average segment diameter (*ASD*)	$ASD(v) = \#vessel\ pixels/L_c(v)$
Distance to disc center (*DDC*)	$DDC(v) = \boldsymbol{c}(a) - \boldsymbol{\rho}.$
Integrated curvature (*IC*)	$IC(v) = \int_a^b \lvert\kappa(s)\rvert ds$
Integrated squared curvature (*ISC*)	$ISC(v) = \int_a^b \lvert\kappa(s)\rvert^2 ds$
Integrated curvature normalized by curve length (*ICLc*)	$ICLc(v) = IC(v)/L_c(v)$
Integrated curvature normalized by chord length (*ICLx*)	$ICLx(v) = IC(v)/(L_x(v))$
Integrated squared curvature normalized by curve length (*ISCLc*)	$ISCLc(v) = ISC(v)/(L_c(v))$
Integrated squared curvature normalized by chord length (*ISCLx*)	$ISCLx(v) = ISC(v)/(L_x(v))$

For *DDC*, $\boldsymbol{c}(a)$ is the start point of segment v, and $\boldsymbol{\rho}$ is the optic disc center. $L_c(v)$ and $L_x(v)$ are the curve length and chord length of segment v, respectively.

Vessel Tracing

The output image of segmentation module is an image with the probability of being a vessel for each pixel. The pixel values change between 0 and 1. Before feeding this image to the feature extraction module, a threshold is applied by using the method of Otsu,[22] which uses histogram information of the image and minimizes the intra-class variance. After applying a threshold, we obtain a vessel image. From this vessel image, we find the center line of vessels and their connectivity information to be able to extract the vessel tree information. We follow the vessel tracing method of Ataer-Cansizoglu et al.,[16,19] and we provide the details of their method in the Supplementary Materials.

Feature Extraction

In image-based machine learning applications, representing an image with informative features plays an important role. Vessel dilation and tortuosity are commonly used for the definition of ROP.[4,5] Also, experts mention that vessel tree information is informative for the diagnosis of ROP.[23,24] Thus, in this part of the system, we aim to extract tortuosity- and dilation-related features from the provided vessel tree information. We represent each image with 11 different feature sets, representing a subset of features that Ataer-Cansizoglu[16] extracted. The features can be divided into two categories: point based or segment based. In addition to the segment-based features presented in Table 1, we extract curvature and point diameter as point-based features. Mathematical derivations of the extracted features are provided in the Supplementary Materials.

The output of the feature extraction stage is a feature vector of size 143: 11 feature sets each containing 13 statistics, 5 based on a Gaussian mixture model fit and 8 on other typical measures.

Classification

We considered a dataset D containing N images, indexed by $i \in \{1, 2, \ldots, N\}$. This dataset contains the tuples of the form $(\boldsymbol{x_i}, y_i)$, where $\boldsymbol{x_i} \in \mathbb{R}^{143}$ is the feature vector generated from Section 3.4, and y_i is the corresponding label of image i. We regressed the image features with their corresponding class labels, and we employed three different classifiers: (1) logistic regression, (2) support vector machine (SVM), and (3) neural networks.

Logistic Regression

We trained a logistic regression classifier with lasso regularization by minimizing the following loss function:

$$L(\beta,\ D) = \sum_{i \in D} \log\left(1 + e^{-y_i \hat{y}_i}\right) + \lambda \|\beta\|_1 \qquad (2)$$

$$\hat{y}_i = \beta^T x_i \qquad (3)$$

where β is the linear-discriminant model parameter, and λ is the regularization weight.

Support Vector Machine

Second, we trained an SVM for classification purposes. We found the learned parameter β by

$$\begin{aligned} &\min\ C \sum_{i \in D} \zeta_i + \beta^T \beta \\ &\text{subject to } y_i\left(\beta^T x_i + b\right) \geq 1 - \zeta_i \\ &\zeta_i \geq 0 \quad \forall i \in D \end{aligned} \qquad (4)$$

Table 2. Classifier Performances Evaluated with Mean AUC (±Confidence Intervals)

			SVM		
		LR	Linear	RBF	NN
Small dataset	Plus vs. not-plus	0.97 (±0.06)	0.95 (±0.08)	0.97 (±0.06)	**0.99 (±0.04)**
	Pre-plus or worse vs. normal	0.98 (±0.03)	0.97 (±0.03)	0.95 (±0.05)	**0.99 (±0.02)**
Large dataset	Plus vs. not-plus	**0.93 (±0.03)**	0.91 (±0.03)	0.92 (±0.03)	0.91 (±0.03)
	Pre-plus or worse vs. normal	**0.87 (±0.02)**	0.86 (±0.02)	0.85 (±0.02)	0.79 (±0.02)

Confidence intervals were calculated from Hanley and McNeil.[25] Classifiers are trained and tested with the features extracted with ground truth optic disc centers.

We used two different kernel functions for training SVMs: linear and radial basis function (RBF).

Neural Network

Finally, we trained a neural network for obtaining a severity score for each image. This network is a fully connected multilayer perceptron. The loss function is the logistic loss used in logistic regression classifier (Equation 2). We tuned the number of layers, number of nodes in each layer, learning rate, and regularization for each dataset and classification task. A detailed explanation for these parameters is provided later.

The regularization parameter λ in logistic regression and neural networks ranged from 10^{-6} to 10^{4}. Also, the penalty term C in SVM ranged from 10^{-2} to 10^{2}. The number of layers in neural networks ranged from 1 to 4, and the number of nodes in the first layer ranged from 1 to 143. The number of nodes in the hidden layers was found by taking the integer part of $n_1/\exp(l)$, where n_1 is the number of nodes in the first layer and l is the depth of the layer. The output layer of the neural networks is a single neuron. We present the best results achieved after tuning these parameters.

Statistical Analysis

In the experiments of testing the severity scores, we binarized labels by using plus versus not-plus and normal versus pre-plus or worse. We calculated the AUC scores with fivefold cross-validation to evaluate the performance of the system. We divided the datasets into five splits such that each split had a near-equal number of samples from each class and represented 20% of the data. Also, the folds do not include any overlapping images of the same patient (acquired in multiple sessions) in the training and test set. We trained each classifier with four splits and tested it on the remaining split. We present the mean AUC of five folds and calculate the 95% confidence intervals by using the formula of Hanley et al.[25] Also, when comparing the prediction performances of our system with the CNN-based method, we assess the statistical significance of the difference by representing the P value of the one-sided Welch's t-test for unequal variances.[26]

Experiments

We trained the optic disc center detector with 5000 randomly chosen images from the large dataset. We extracted two sets of features on both the small and large datasets by relying on (1) optic disc centers provided by experts, and (2) optic disc centers predicted by the detector we trained. We then trained and tested the classifiers with both sets of features.

Results

We calculated the results of the three classifiers in two methods, using manual identification of the optic disc (Table 2) and with CNN detection (Table 3). As shown in Table 2, the neural networks provided a slightly higher mean AUC in predicting both plus and normal than all the other classifiers for the small dataset. It achieved 0.99 AUC for predicting both plus and normal classes. These results obtained after cross-validation determined that the best neural networks to predict plus and normal classes are two-layer neural networks with (75, 1) and (140, 1) neurons in each layer, respectively. Values of the regularization parameter λ for the networks were 0.001 and 1, respectively, as a result of cross-validation. Neural networks were trained using gradient descent with respective step sizes of 0.005 and 0.001 over 500 epochs. It is also worth noting that, for the large dataset, logistic regression provided higher mean AUCs than other classifiers in predicting both plus and normal classes. Values of the regularization parameter λ for logistic regression classifiers of plus and normal labels were identified in cross-validation as 4 and 1, respectively.

Table 3 shows the results of the second set of experiments, where the optic disc center was found with

Table 3. Classifier Performances Evaluated with Mean AUC (±Confidence Intervals)

			SVM		
		LR	Linear	RBF	NN
Small dataset	Plus vs. not-plus	0.98 (±0.05)	0.92 (±0.10)	0.90 (±0.11)	**0.99 (±0.04)**
	Pre-plus or worse vs. normal	**0.98 (±0.03)**	0.95 (±0.05)	0.97 (±0.03)	0.97 (±0.03)
Large dataset	Plus vs. not-plus	0.92 (±0.03)	0.92 (±0.03)	0.89 (±0.03)	**0.94 (±0.03)**
	Pre-plus or worse vs. normal	0.87 (±0.02)	0.86 (±0.02)	0.85 (±0.02)	**0.88 (±0.01)**

Confidence intervals were calculated from Hanley and McNeil.[25] Classifiers are trained and tested with the features extracted with predicted optic disc centers.

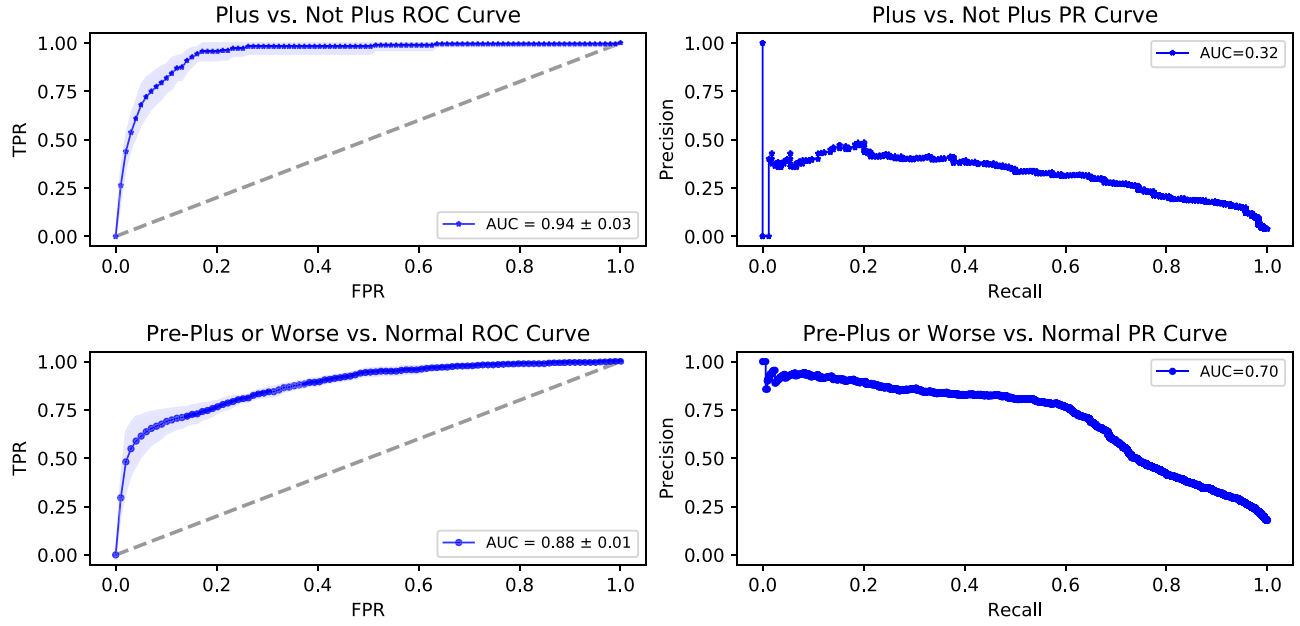

Figure 3. ROC and PR curves of neural networks trained and tested on features of the large dataset with predicted optic disc centers. The plots on the first and second row are the ROC and PR curves of networks predicting plus versus not-plus and pre-plus or worse versus normal, respectively.

the detector described earlier. For the small dataset, the neural networks provided the highest mean AUC compared to the other classifiers in predicting plus with 0.99 AUC. The neural network for predicting the plus labels is a three-layer neural network with (25, 9, 1) neurons at each layer. The regularization parameter λ for training this network was 0.5. We used gradient descent to optimize the network parameters with a step size of 0.01 over 300 epochs. When predicting the normal labels in the small dataset, logistic regression achieved 0.98 AUC, which is slightly higher than the other classifiers. The regularization parameter λ for this classifier during training was 1000. Also, for the large dataset, the neural networks provided a slightly higher mean AUC in predicting both plus and normal labels than all the other classifiers: 0.94 and 0.88, respectively. The neural network models achieved this performance with a two-layer network with (36, 1) neurons in each layer. Values of the regularization parameter λ for training the networks were 0.001 for plus and 1 for normal. Gradient descent with all data was used to optimize the network parameters using a step size of 0.1 over 200 epochs for plus and a step size of 0.01 over 500 epochs for normal. Figure 3 shows the ROC and PR curves of the corresponding networks.

Discussion

We have reported the results of a fully automated feature extraction-based pipeline for plus disease classification. There are several key findings. These results suggest that the feature-extraction-based approach, I-ROP ASSIST, achieves high performance in predicting plus disease, and this performance is similar to the

performance of CNN-based approaches. Next, multiple classifiers produce high or similar performance, suggesting that the extracted features are meaningful and the success-limiting step of feature-based work. Finally, we also claim that extracted features are linearly separable.

As noted, the first key finding is that CNN-like performance can be achieved with feature-extraction-based methods. The CNN approach developed by Brown et al.[13] achieved 0.94 and 0.98 AUC statistics for predicting pre-plus or worse versus normal and plus versus not-plus disease, respectively. Our feature-based approach achieved 0.88 and 0.94 AUC values for the corresponding tasks. The AUC differences between the models are 0.6 and 0.4. The *P* values corresponding to these compressions are 0.88 and 0.99, respectively. These results suggest that the CNN could identify more discriminative features given a large dataset for training, whereas the curated features we designed have been a performance-limiting factor, resulting in approximately 5% lower AUC.

The second and third key findings stem from having near-equal results from different classifiers. Achieving high performance using the feature set described earlier in the Feature Extraction section shows that the extracted features are discriminative for ROP classification. Having near-equal results from classifiers that search for both linear and nonlinear boundaries among classes indicates that, instead of the separation of the features, the extracted features are the main determining factor for our model's classification accuracy levels. Moreover, the fact that the linear logistic classifier performance was competitive with nonlinear alternatives such as the RBF–SVM and NNs indicates that the optimal classification boundaries for the given feature distributions are close to being linear. This indicates that even very simple classifiers such as the logistic classifier can achieve high performance when classifying the handcrafted features.

CNN methods have demonstrated significant advances in medical problems; however, their uninterpretable black-box nature is a barrier to real-life applications. Clinicians are often unwilling to accept their recommendations without explanation, which is currently not provided even by state-of-the-art CNN methods.[27–30] However, the I-ROP ASSIST system uses only handcrafted features relevant to the definition of ROP.

Limitations

The quality of the input images is a potential limitation for the I-ROP ASSIST system. The system requires 480 × 640-pixel input images and resizes the input images with different shapes before processing. Because resizing the image will affect the image quality, downscaling larger images and upscaling smaller images are other potential limitations of this study. Also, studies show that inconsistency for the clinical diagnosis of plus disease is significant even among experts.[31,32] To mitigate this potential limitation in this study, we used reference standard diagnoses.[17]

Future Work

The front-end image processing stage offers a rich environment in which better features can be discovered. The statistics we used as features are quite coarse, considering the level of detail and potentially discriminative information present in the segmented and traced vessels. In future work, we will improve the feature extraction process to achieve better performance levels. Also, we will extend our work to a system that uses our handcrafted features for explaining CNN predictions. We showed that both CNN and I-ROP ASSIST extract relevant features for ROP diagnosis. Because of the black-box nature of CNN, the CNN features are not explainable. We will use our handcrafted features for finding a mapping between CNN features and features that clinicians can understand.

Conclusions

This fully automated system, which combines retinal vessel segmentation, tracing, feature extraction and classification stages, diagnosed plus disease in ROP with performance on par with recent publications reporting on the use of CNNs. Combining these approaches in the future may lead to improved explainability of deep CNNs. The MIT-licensed complete code package is available to the public at https://github.com/neu-spiral/iROPASSISTPackage. We also provide the features of three example images from our dataset for public use.

Acknowledgments

The Imaging and Informatics in Retinopathy of Prematurity (i-ROP) Research Consortium includes the following: Michael F. Chiang, Susan Ostmo,

Sang Jin Kim, Kemal Sonmez, and J. Peter Campbell (Oregon Health & Science University, Portland, OR); R. V. Paul Chan and Karyn Jonas (University of Illinois at Chicago, Chicago, IL); Jason Horowitz, Osode Coki, Cheryl-Ann Eccles, and Leora Sarna (Columbia University, New York, NY); Anton Orlin (Weill Cornell Medical College, New York, NY); Audina Berrocal and Catherin Negron (Bascom Palmer Eye Institute, Miami, FL); Kimberly Denser, Kristi Cumming, Tammy Osentoski, Tammy Check, and Mary Zajechowski (William Beaumont Hospital, Royal Oak, MI); Thomas Lee, Evan Kruger, and Kathryn McGovern (Children's Hospital Los Angeles, Los Angeles, CA); Charles Simmons, Raghu Murthy, and Sharon Galvis (Cedars Sinai Hospital, Los Angeles, CA); Jerome Rotter, Ida Chen, Xiaohui Li, Kent Taylor, and Kaye Roll (Los Angeles Biomedical Research Institute, Los Angeles, CA); Jayashree Kalpathy-Cramer (Massachusetts General Hospital, Boston, MA); Deniz Erdogmus and Stratis Ioannidis (Northeastern University, Boston, MA); Maria Ana Martinez-Castellanos, Samantha Salinas-Longoria, Rafael Romero, Andrea Arriola, Francisco Olguin-Manriquez, Miroslava Meraz-Gutierrez, Carlos M. Dulanto-Reinoso, and Cristina Montero-Mendoza (Asociacion para Evitar la Ceguera en Mexico, Mexico City, Mexico).

This project was supported by grants R01EY19474, K12EY027720, and P30EY10572 from the National Institutes of Health; by grants SCH-1622679, SCH-1622542, and SCH-1622536 from the National Science Foundation; and by unrestricted departmental funding and a Career Development Award from Research to Prevent Blindness (JPC).

Disclosure: **V.M. Yildiz**, None; **P. Tian**, None; **I. Yildiz**, None; **J.M. Brown**, None; **J. Kalpathy-Cramer**, None; **J. Dy**, None; **S. Ioannidis**, None; **D. Erdogmus**, None; **S. Ostmo**, None; **S.J. Kim**, None; **R.V.P. Chan**, Visunex Medical Systems Scientific Advisory Board (S), Genentech (C); **J.P. Campbell**, None; **M.F. Chiang**, Clarity Medical Systems Scientific Advisory Board (S), Novartis (C), and Inteleretina (I)

References

1. Fierson WM, American Academy of Pediatrics Section on Ophthalmology, American Academy of Ophthalmology, American Association for Pediatric Ophthalmology and Strabismus and American Association of Certified Orthoptists. Screening examination of premature infants for retinopathy of prematurity. *Pediatrics*. 2013;131:189–195.
2. Gilbert C, Foster A. Childhood blindness in the context of VISION 2020: the right to sight. *Bull World Health Org*. 2001;79:227–232.
3. Gilbert C, Fielder A, Gordiollo L, et al. Characteristics of infants with severe retinopathy of prematurity in countries with low, moderate, and high levels of development: implications for screening programs. *Pediatrics*. 2005;115:518–525.
4. International Committee for the Classification of Retinopathy of Prematurity. An international classification of retinopathy of prematurity. *Arch Ophthalmol*. 1984;102:1130–1134.
5. International Committee for the Classification of Retinopathy of Prematurity. The International Classification of Retinopathy of Prematurity revisited. *Arch Ophthalmol*. 2005;123:991–999.
6. Clare MW, Kenneth DC, Merrick JM, et al. Computerized analysis of retinal vessel width and tortuosity in premature infants. *Invest Ophthalmol Vis Sci*. 2008;49:3577–3585.
7. Gelman R, Martinez-Perez ME, Vanderveen DK, Muskowitz A, Fulton AB. Diagnosis of plus disease in retinopathy of prematurity using retinal image multiscale analysis. *Invest Ophthalmol Vis Sci*. 2005;46:4734–4738.
8. Tsai CL, Madore B, Leotta MJ, et al. Automated retinal image analysis over the internet. *IEEE Trans Inf Technol Biomed*. 2008;12:480–487.
9. Fiorin D, Ruggeri A. Computerized analysis of narrow-field ROP images for the assessment of vessel caliber and tortuosity. *Conf Proc IEEE Eng Med Biol Soc*. 2011;2011:2622–2625.
10. Pour EK, Pourreza H, Zamani KA, et al. Retinopathy of prematurity-assist: novel software for detecting plus disease. *Korean J Ophthalmol*. 2017;31:524–532.
11. Wallace DK, Jomier J, Aylward SR, Landers MB, 3rd. Computer-automated quantification of plus disease in retinopathy of prematurity. *J AAPOS*. 2003:7:126–130.
12. Brown JM, Campbell JP, Beers A, et al. Fully automated disease severity assessment and treatment monitoring in retinopathy of prematurity using deep learning. In: *Proceedings of SPIE 10579, Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*. 2018;105790Q.
13. Brown JM, Campbell JP, Beers A, et al. Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. *JAMA Ophthalmol*. 2018;136:803–810.

14. Worrall DE, Wilson CM, Gabriel JB. Automated retinopathy of prematurity case detection with convolutional neural networks. In: Carneiro G, Mateus D, Loïc P, et al., eds. *Deep Learning and Data Labeling for Medical Applications*. Cham, Switzerland: Springer International; 2016:68–76.
15. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60–88.
16. Ataer-Cansizoglu E. *Retinal Image Analytics: A Complete Framework from Segmentation to Diagnosis*. Boston, MA: Northeastern University; 2015.
17. Ryan MC, Ostmo S, Jonas K, et al. Development and evaluation of reference standards or image-based telemedicine diagnosis and clinical research studies in ophthalmology. *AMIA Annu Symp*. 2014;2014:1902–1910.
18. Kalpathy-Cramer J, Campbell JP, Erdogmus D, et al. Plus disease in retinopathy of prematurity: improving diagnosis by ranking disease severity and using quantitative image analysis. *Ophthalmology*. 2016;123:2345–2351.
19. Ataer-Cansizoglu E, Bolon-Canedo V, Campbell JP, et al. Computer-based image analysis for plus disease diagnosis in retinopath of prematurity: performance of the "i-ROP" system and image features associated with expert diagnosis. *Transl Vis Sci Technol*. 2015;4:5.
20. Campbell JP, Kalpathy-Cramer J, Erdogmus D, et al. Plus disease in retinopathy of prematurity: a continuous spectrum of vascular abnormality as basis of diagnostic variability. *Ophthalmology*. 2016;123:2338–2344.
21. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, eds. *Medical Image Computing and Computer-Assisted Intervention*. Cham, Switzerland: Springer International; 2015:234–241.
22. Otsu N. A threshold selection method from gray level histograms. *IEEE Trans Syst Man Cybern*. 1979;9:62–66.
23. Rao R, Jonsson NJ, Ventura C, et al. Plus disease in retinopathy of prematurity: diagnostic impact of field of view. *Retina*. 2012;32:1148–1155.
24. Thyparampil PJ, Park Y, Martinez-Perez ME, et al. Plus disease in retinopathy of prematurity: quantitative analysis of vascular change. *Am J Ophthalmol*. 2010;150:468–475.
25. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143:29–36.
26. Sawilowsky SS. Fermat, Schubert, Einstein, and Behrens-Fisher: the probable difference between two means when $\sigma 1^2 \neq \sigma 2^2$. *J Mod Appl Statist Meth*. 2002;1:461–472.
27. Castelvecchi D. Can we open the black box of AI? *Nature*. 2016;538:20–23.
28. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. *JAMA*. 2018;320:2199–2200.
29. Petkovic D, Kobzik L, Re C. Machine learning and deep analytics for biocomputing: call for better explainability. *Pac Symp Biocomput*. 2018;23:623–627.
30. Ting DSW, Peng L, Varadarajan AV, et al. Deep learning in ophthalmology: the technical and clinical considerations. *Prog Retin Eye Res*. 2019;72:100759.
31. Chiang MF, Jiang L, Gelman R, Du YE, Flynn JT. Interexpert agreement of plus disease diagnosis in retinopathy of prematurity. *Arch Ophthalmol*. 2007;125:875–880.
32. Wallace DK, Quinn GE, Feedman SF, Chiang MF. Agreement among pediatric ophthalmologists in diagnosing plus and pre-plus disease in retinopathy of prematurity. *J AAPOS*. 2008;12:352–356.
33. Hastie T, Stuetzle W. Principal curves. *J Am Stat Assoc*. 1989;84:502–516.
34. Ozertem U, Erdogmus D. Locally defined principal curves and surfaces. *J Mach Learn Res*. 2011;12:1249–1286.
35. Erdogmus D, Ozertem U. Self-consistent locally defined principal surfaces. *IEEE Int Conf Acoustics, Speech, Signal Process*. 2007;2:549–552.
36. Shaker M, Myhre JN, Erdogmus D. Computationally efficient exact calculation of kernel density derivatives. *J Signal Process Syst*. 2015;81:321–332.