Automated Image Quality Assessment for Fundus Images in Retinopathy of Prematurity

Aaron S. Coyner[1,2], Ryan Swan[1,2], Jayashree Kalpathy-Cramer[3], Sang Jin Kim[1,4], J. Peter Campbell[1], Karyn E. Jonas[5], Susan Ostmo[1], R.V. Paul Chan[6], Michael F. Chiang[1]

[1]Ophthalmology, [2]Medical Informatics, Oregon Health & Science University, Portland, OR United States.
[3]MGH/Harvard Medical School, Boston, MA, United States.
[4]Ophthalmology, Sungkyunkwan University School of Medicine, Seoul, Korea.
[5]University of Illinois at Chicago, Chicago, IL, United States.
[6]Ophthalmology, Illinois Eye and Ear Infirmary, Chicago, IL, United States.

PURPOSE

Accurate image-based ophthalmic diagnosis relies on clarity of fundus images.[1] This has important implications for the quality of ophthalmic diagnosis, and for emerging methods such as telemedicine and computer-based image analysis.[1-2] Convolutional neural nets (CNN) have continued to show promise in image recognition tasks[3]; therefore, the aim of this study was to implement a CNN for automatically assessing the quality of fundus images, and to evaluate its performance on a set of Retinopathy of Prematurity (ROP) images compared to expert assessment.

METHODS

IMAGE COLLECTION

A set of 5,174 wide-angle fundus images (RetCam; Natus, Pleasanton, CA) was collected from preterm infants with a possible diagnosis of ROP. In addition to many other metrics, images were assessed for quality and whether or not an accurate ROP diagnosis could be made. Possible quality grades were: “Acceptable for diagnosis,” “Possibly acceptable for diagnosis,” and “Not acceptable for diagnosis.” Because there were few images labeled “Not acceptable for diagnosis” (~200), images labeled “Possibly acceptable for diagnosis” and “Not acceptable for diagnosis” were placed into a single group. To simplify nomenclature, images in the “Acceptable for diagnosis” set were labeled “optimal” and images from the combined group of “Not acceptable for diagnosis” and “Possibly acceptable for diagnosis” were labeled “suboptimal.”

IMAGE SETS

2,550 images were used for training, 1,000 images were used for validation, and 1,624 images were used as a test set. Training and validation sets were equi-representative of each class, and the training set resembled the underlying distribution of each class (797 optimal : 827 suboptimal).

In addition, 30 images of varying quality were selected for review by six individual expert graders. Using an online Elo rating system, graders performed pairwise comparisons between images until all images were ranked from worst quality to best quality. A consensus rank for the image set was developed from individual ranks.

CONVOLUTIONAL NEURAL NET

A convolutional neural net (CNN) was implemented in Python using the Keras package with a TensorFlow backend (Figure 1). The validation set error and loss were used to monitor the training after each epoch. When the validation error began to level off/increase, training was discontinued. The weights of the CNN were saved and used for assessing the test set. The smaller set of 30 images of varying quality were also assessed and, using the probability of each image belonging to the optimal image class, the image set was ranked from worst to best quality.

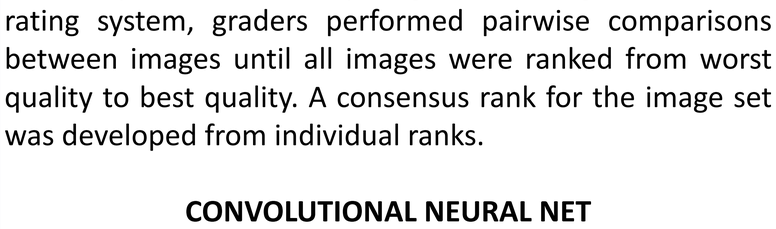

Figure 1: CNN Architecture
This project used an adapted version of AlexNet.[3] Images were down-sampled to 32x32 RGB for input, sent through three convolutional layers, and two dense layers.

ACKNOWLEDGEMENTS

This work is supported by National Library of Medicine Training Grant 4T15LM007088-25, National Institutes of Health (R01EY019474, P30EY10572, P41EB015896), by the National Science Foundation (SCH-1622542, SCH-1622536, SCH-1622679), and by unrestricted departmental funding from Research to Prevent Blindness.

RESULTS

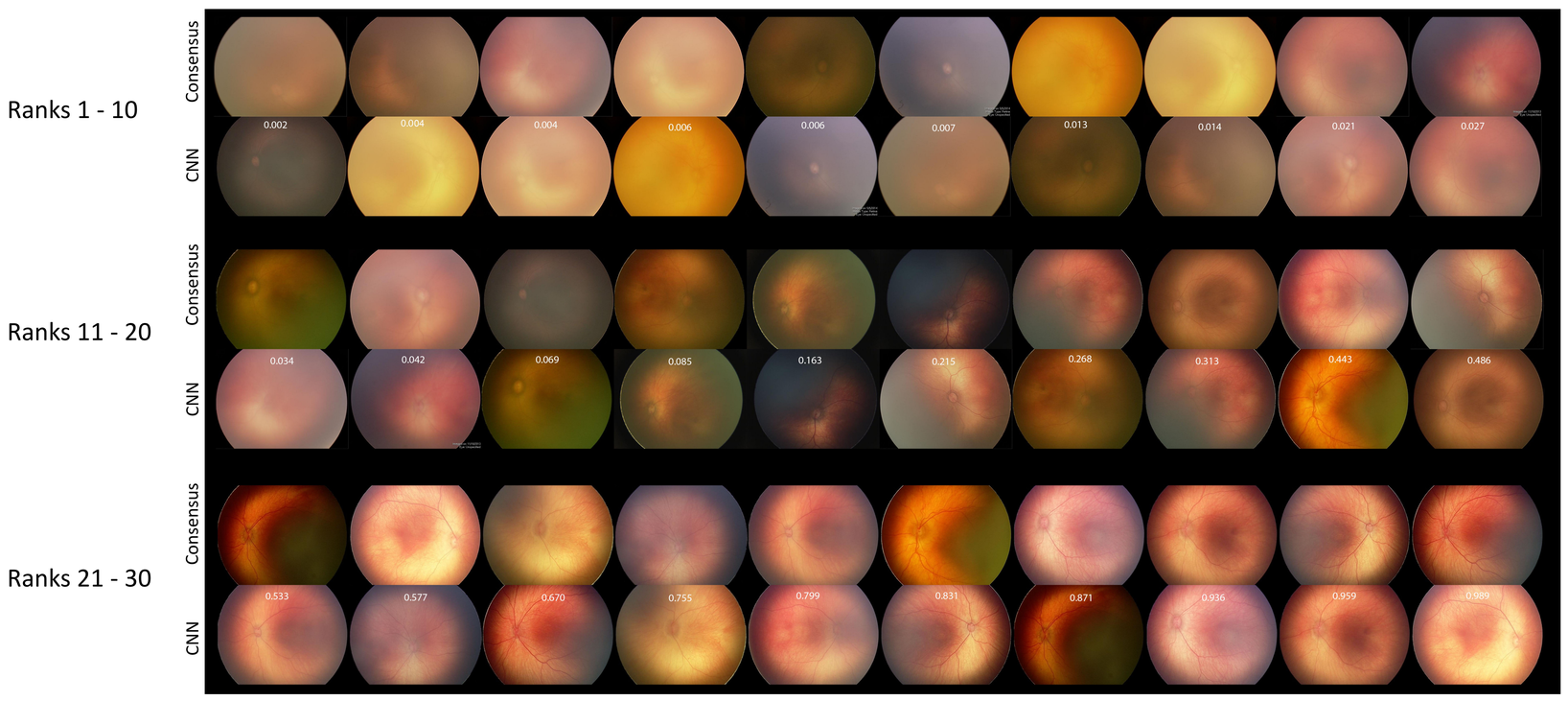

Figure 2: Montage of Ranked Images from Expert Consensus and Algorithm
Images, ranked 1 (worst quality) through 30 (best quality), are displayed in three rows. Each row shows the image in its corresponding rank for the expert consensus rank (top sub-row) and the algorithm rank (bottom sub-row). Qualitatively, the algorithm is able to separate images of lower quality from images of higher quality. Red box illustrates a discrepancy between the algorithm rank and the consensus rank.

	CNN	Consensus	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Expert 6
CNN	1.00	0.82	0.78	0.82	0.82	0.83	0.84	0.85
Consensus	0.82	1.00	0.98	0.94	0.94	0.97	0.96	0.96
Expert 1	0.78	0.98	1.00	0.90	0.89	0.94	0.93	0.93
Expert 2	0.82	0.94	0.90	1.00	0.94	0.96	0.92	0.95
Expert 3	0.82	0.94	0.89	0.94	1.00	0.91	0.91	0.92
Expert 4	0.83	0.97	0.94	0.96	0.91	1.00	0.94	0.97
Expert 5	0.84	0.96	0.93	0.92	0.91	0.94	1.00	0.97
Expert 6	0.85	0.96	0.93	0.95	0.92	0.97	0.97	1.00

Table 1: Correlation Matrix of Expert Grader Rank and Algorithm Rank
Spearman's rank-order correlation matrix of CNN image rank, expert consensus rank, and ranks from the six individual experts.

True Label

Predicted Label	Suboptimal	Optimal
Suboptimal	803	65
Optimal	24	732

Table 2: Confusion Matrix
Confusion matrix derived from test set. Accuracy of CNN on test set is 94.5%.

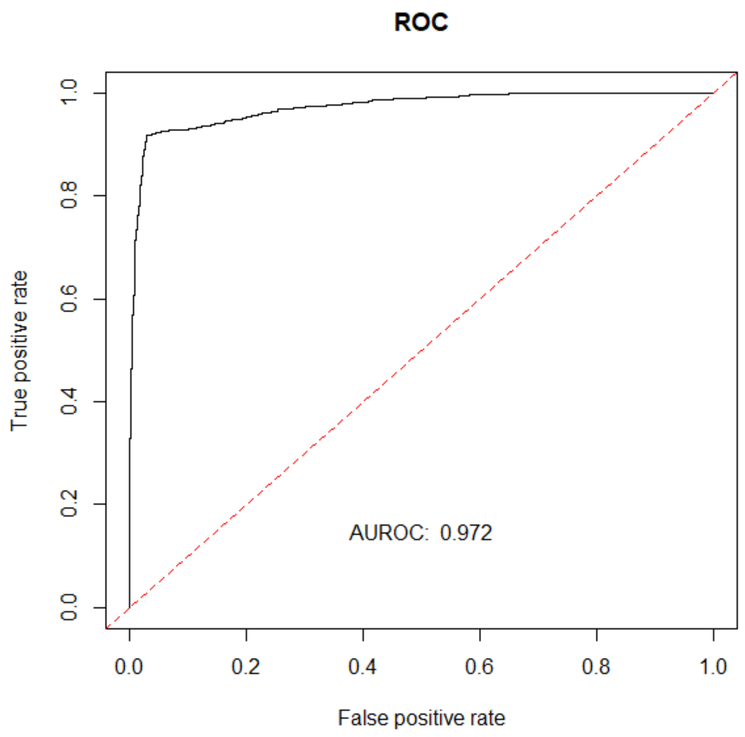

Figure 3: Receiver Operating Characteristics Curve
This plot shows the receiver operating characteristics (ROC) curve. The area under the ROC curve (AUROC) is 0.972, indicating that the model can discriminate between optimal and suboptimal images well.

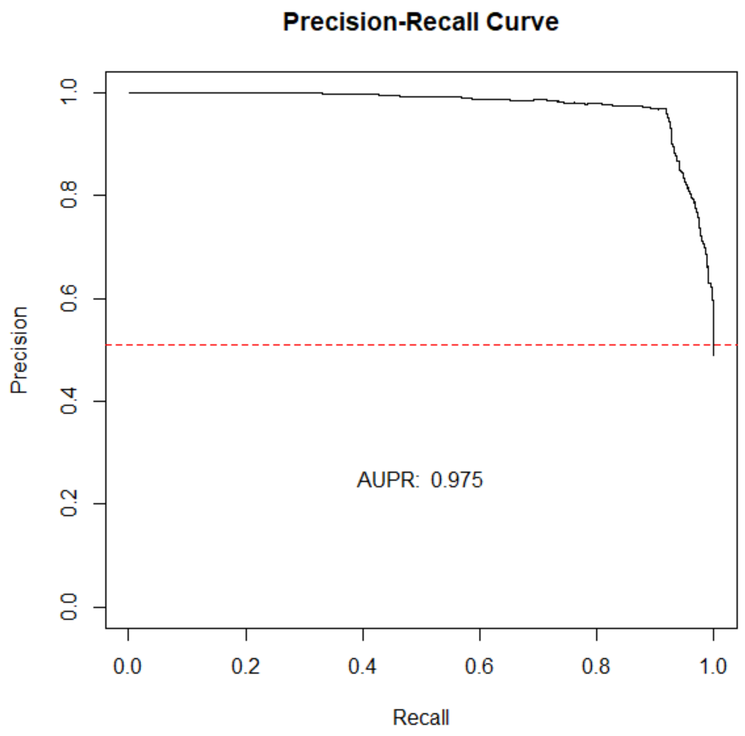

Figure 4: Precision-Recall Curve
This plot shows the precision-recall curve. The area under the this curve (AUPRC) is 0.975. This indicates that the model has both a low false positive rate and a low false negative rate.

CONCLUSIONS

- Expert grader image rankings are highly correlated with one another (correlation coefficient [CC] 0.89-0.94) and with the consensus ranking (CC 0.94 – 0.98).
- CNN image ranking is moderately correlated with the consensus ranking (CC 0.82) and individual expert rankings (CC 0.78- 0.85).
- The CNN can reliably distinguish optimal from suboptimal images. On the validation set, the accuracy was 95.1%. On the test set, accuracy was 94.5%, area under the receiver operating curve (AUROC) was 0.972, and area under the precision-recall curve (AUPR) was 0.975.

REFERENCES

1. Chiang MF, Wang L, Busuioc M, Du YE, Chan P, Kane SA, Lee TC, Weissgold DJ, Berrocal AM, Coki O, Flynn JT, Starren J. Telemedical Retinopathy of Prematurity Diagnosis Accuracy, Reliability, and Image Quality. Arch Ophthalmol. 2007;125(11):1531-1538. doi:10.1001/archopht.125.11.1531
2. Richter GM, Williams SL, Starren J, Flynn JT, Chiang MF. Telemedicine for Retinopathy of Prematurity Diagnosis: Evaluation and Challenges. Survey of ophthalmology. 2009;54(6):671-685. doi:10.1016/j.survophthal.2009.02.020.
3. Krizhevsky A, Sutskever I, and Hinton G. (2012), ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems 25, 1097-1105