AMERICAN ACADEMY OF OPHTHALMOLOGY®

Improved Training Efficiency for Retinopathy of Prematurity Deep Learning Models Using Comparison versus Class Labels

Adam Hanif, MD,[1] İlkay Yıldız, PhD,[2] Peng Tian, PhD,[2] Beyza Kalkanlı, BS,[2] Deniz Erdoğmuş, PhD,[2] Stratis Ioannidis, PhD,[2] Jennifer Dy, PhD,[2] Jayashree Kalpathy-Cramer, PhD,[3] Susan Ostmo, MS,[1] Karyn Jonas, BSN,[4] R. V. Paul Chan, MD, MBA,[4] Michael F. Chiang, MD,[5] J. Peter Campbell, MD, MPH[1]

Purpose: To compare the efficacy and efficiency of training neural networks for medical image classification using comparison labels indicating relative disease severity versus diagnostic class labels from a retinopathy of prematurity (ROP) image dataset.

Design: Evaluation of diagnostic test or technology.

Participants: Deep learning neural networks trained on expert-labeled wide-angle retinal images obtained from patients undergoing diagnostic ROP examinations obtained as part of the Imaging and Informatics in ROP (i-ROP) cohort study.

Methods: Neural networks were trained with either class or comparison labels indicating plus disease severity in ROP retinal fundus images from 2 datasets. After training and validation, all networks underwent evaluation using a separate test dataset in 1 of 2 binary classification tasks: normal versus abnormal or plus versus nonplus.

Main Outcome Measures: Area under the receiver operating characteristic curve (AUC) values were measured to assess network performance.

Results: Given the same number of labels, neural networks learned more efficiently by comparison, generating significantly higher AUCs in both classification tasks across both datasets. Similarly, given the same number of images, comparison learning developed networks with significantly higher AUCs across both classification tasks in 1 of 2 datasets. The difference in efficiency and accuracy between models trained on either label type decreased as the size of the training set increased.

Conclusions: Comparison labels individually are more informative and more abundant per sample than class labels. These findings indicate a potential means of overcoming the common obstacle of data variability and scarcity when training neural networks for medical image classification tasks. *Ophthalmology Science 2022;2:100122 © 2022 by the American Academy of Ophthalmology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).*

A deep learning model's performance is associated strongly with the volume and quality of data on which it has been trained.[1–5] In use cases involving medical image classification, datasets traditionally comprise images with human-assigned class labels indicating the represented diagnosis or finding. However, large, well-composed datasets containing high-quality images for these purposes are not always feasible to obtain.[1] Furthermore, the process of acquiring such images and enlisting the help of expert graders to assign labels is both labor intensive and prone to high interlabeler variance. An alternative method for training deep learning models has been described using comparison labels obtained from human-drawn comparisons of 2 data inputs in the set.[6,7] In the context of medical diagnosis, this may involve experts grading the relative severity of disease in multiple pairwise comparisons of cases within a dataset and assigning labels to indicate the ranking. This approach has the added advantage of assessing disease severity along a continuum, rather than in categories, which may reflect the natural distribution of disease phenotypes more accurately.

Incorporation of comparison labels offers 2 theoretical advantages in the training process. First, training with labels representing all possible relative comparisons between each input in the dataset increases the number of labels for use in training quadratically, potentially improving the performance of models trained on smaller datasets. Second, grading of disease through comparison of severity has demonstrated less intergrader variability than classification alone, suggesting that the creation of a training set with less noise than other labeling methods.[8–10] Although potentially more labor intensive to obtain, the method may lead to more accurate and efficient neural network training with limited amounts of data. In this project, we applied this concept to explore the relative efficiency of neural networks trained to predict disease severity using comparison labels versus the traditional method of using diagnostic class labels on a retinopathy of prematurity (ROP) image dataset.

https://doi.org/10.1016/j.xops.2022.100122
ISSN 2666-9145/22

Methods

Neural Network

A neural network architecture inspired by Siamese networks was constructed using Bradley-Terry and Thurstone models as loss functions and designed to learn from class and comparison labels as described previously.[11,12] This network expands on the conventional application of Siamese networks not only by predicting the similarity between inputs of 2 identical base networks, but also by regressing comparison labels simultaneously.[13] When training with comparison labels, the network learns by maximizing the likelihood of comparison labels under the Bradley-Terry model.[11,12] When learning from class labels, the network uses the same architecture as a base Siamese network, predicting the class label pertaining to the single input image.

Formally, a base neural network exists representing the coupling between class and comparison labels. The base neural network receives an image and produces latent features that are predictive of both class and comparison labels. The classification network contains the base network, followed by a fully connected neural network that predicts the class label from latent features extracted by the base network. The comparison network receives a pair of images and extracts the corresponding pair of latent features using the same base network. The base network is followed by another fully connected neural network that predicts the severity score from the latent features of each image. Finally, the pair of severity scores are used collectively to predict the comparison label outcome between the pair of images.

Datasets

Three pre-existing datasets were used in the study comprising wide-angle retinal images obtained from patients undergoing diagnostic ROP examinations with digital fundus imaging using the RetCam (Natus Medical, Inc). All images exhibited the posterior retina and were obtained as part of the Imaging and Informatics in ROP (i-ROP) cohort study. Two labeled datasets were used to train the network using class and comparison labels. The first dataset included 100 retinal images labeled by members of the i-ROP consortium (the i-ROP dataset). The second dataset included 30 images labeled by the 34 members of the Third International Classification of Retinopathy of Prematurity (ICROP) committee (the ICROP dataset). A test dataset comprising 5561 separate retinal images was used for evaluation of the classification and comparison neural networks (Table 1). All images in the i-ROP and test datasets were assigned a reference standard diagnosis based on the consensus diagnosis among 3 masked image graders and the ophthalmoscopic diagnosis, as described previously.[14] This study was approved by the institutional review board at Oregon Health & Science University and all participating institutions (Beaumont Health, Cedars Sinai Medical Center, Children's Hospital of Los Angeles, Columbia University Medical Center, Weill Cornell Medical Center, University of Miami Health System) in the i-ROP cohort study. The research adhered to the tenets of the Declaration of Helsinki, and written informed consent was obtained from all parents of infants whose images were included in the datasets.

Labeling

The i-ROP and ICROP datasets were labeled in 2 ways: with a class label for each image and with a comparison label for each pair of images. For both datasets used in training and validation, classification and comparison were performed using an open-source, web-based, image severity assessment platform as described previously.[10] For class labels, each grader first was presented 1 image at a time and was asked to assign each image a label of no plus, preplus, and plus (Fig 1). Next, for comparison labels, graders were presented a pair of images from within the dataset and prompted to click on the image that represents more severe disease. In instances where images of similar severity were presented, reviewers were expected to choose which was of marginally greater severity based on their best clinical intuition and experience. The Elo algorithm was used to convert pairwise comparisons from this task into rankings.[15] One "class label" indicating severity of disease was assigned per image per grader. For the i-ROP dataset, 13 expert graders were recruited to provide class labels, and 5 experts completed the comparison task (consisting of more than 4000 pairwise comparisons among the 100 images). The 30 images in the ICROP dataset were labeled both at the image level (class) and by pairwise comparisons by the 34 members of the Third ICROP committee.

Table 1. Distribution of Plus Disease Severity Classes within Datasets

Dataset	Normal	Preplus	Plus	Total
i-ROP	54	31	15	100
ICROP	6	10	14	30
Test dataset	4577	812	172	5561

ICROP = International Classification of Retinopathy of Prematurity; i-ROP = Imaging and Informatics in ROP.

Experimental Design

Both the i-ROP and ICROP datasets were used separately for training of neural networks in 2 primary experiments. Twenty images with reference standard diagnosis labels from the i-ROP dataset (that were not used in training or testing) were used to optimize each trained model for the classification task in a validation step. The test dataset then was used to measure the performance of the best class and comparison models from each dataset. This training, validation, and testing scheme (Fig 2) was performed with incrementally smaller training sets comprising either class or comparison labels corresponding to a fixed number of randomly selected images within the dataset (experiment A) or a fixed number of randomly selected labels (experiment B). Each experiment was performed 3 times, each generating an area under the receiver operating characteristic curve (AUC) per training set size in 1 of 2 binary classification tasks: normal versus abnormal (preplus and plus) and plus versus nonplus (normal and preplus). For each of the 3 repetitions at each size of training set, a different group of randomly selected images from within the corresponding image dataset was used.

From each dataset, 60% of the available images and their corresponding labels given by the number of experts, E, were selected randomly for use in a training subset (Fig 3). This subset then was refined to compose a balanced distribution of the 3 possible severity classes. First, 1 expert whose gradings within the subset were distributed across these classes most evenly compared with other graders was identified. The number of labels for the severity class assigned least frequently by this grader then was determined and was used as the number of images to sample randomly from each class type. The resulting sum, N, of these selected images and their corresponding class and comparison labels, M, collected from all experts constituted the final balanced training set, comprising equal numbers of images of each severity class. This allowed the use of the largest samples

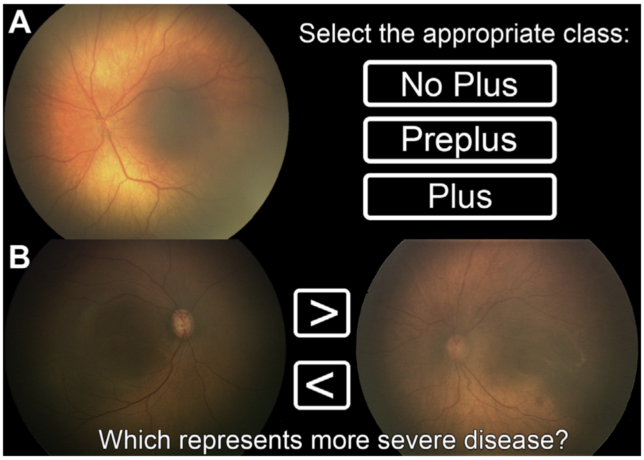
A
Select the appropriate class:
No Plus
Preplus
Plus
B
>
<
Which represents more severe disease?

Figure 1. Diagram showing the labeling process. Graders were asked to perform 2 tasks. **A**, They were given a single image at a time and asked to label the image as plus, preplus, or no plus. **B**, They were shown a pair of images and asked to choose the image that represented more severe disease.

possible, because the limiting batch size was the minority class of the labeler with the largest number in the minority class.

In experiment A, neural networks were trained on either $N \times E$ class labels corresponding to the images within the final balanced training set, or all comparison labels corresponding to the same set of images. In experiment B, neural networks were trained on either M / E images with M class labels within the final balanced training set, or M / E randomly selected image pairs with M comparison labels associated with the same set of images. In later iterations of both experiments, the number of images (N) or labels (M) used in training was reduced incrementally.

In experiment A, neural networks were trained on either N (total selected images) $\times$ E (number of graders who assigned class labels to the images in N) class labels corresponding to the images within the final balanced training set, or all comparison labels corresponding to the same set of images. In experiment B, neural networks were trained on either M / E images with M class labels within the final balanced training set, or M / E randomly selected image pairs with M comparison labels associated with the same set of images. In later iterations of both experiments, the number of images (N) or labels (M) used in training was reduced incrementally.

Neural Network Implementation

The training procedure follows closely from Yıldız et al[6] and Brown et al.[16] Retinal images are prepared first with a pretrained U-Net architecture to convert the colored images into black-and-white masks for retinal vessels.[17] The GoogleNet convolutional neural network architecture,[18] without the fully connected layers, is used as the base neural network to extract latent features from each image. As required by the GoogleNet architecture design, each image is resized to 224 × 224. To leverage the well-known transfer learning properties of neural networks trained on images, GoogleNet layers are initialized with weights pretrained on the ImageNet dataset.[19] Both fully connected networks following the base network in classification and comparison networks are designed as single fully connected layers with sigmoid activations. Classification and comparison networks are trained separately end to end via stochastic gradient descent, in which the learning rate is varied in the range 0.01 to 0.0001. To avoid overfitting when learning from a small number of training images, weight decay is used with regularization parameter varying in the range 0.02 to 0.0002. Both learning rate and regularization hyperparameters are selected with respect to the prediction performance on the validation set.

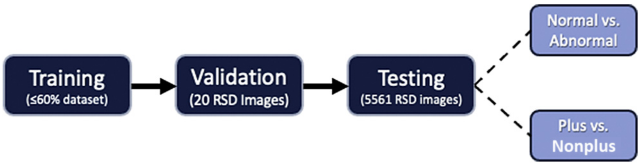

Figure 2. Schematic diagram illustrating the training, validation, and testing process involved in developing the neural networks applied to 1 of 2 binary classification tasks: normal versus abnormal and plus versus nonplus. RSD = reference standard diagnosis.

Having learned from comparison labels via the comparison network, the severity score predicted for each image can be used for both class and comparison predictions. Comparison label prediction follows the same procedure as training, in which a pair of severity scores extracted from a pair of images are used collectively to predict the comparison label. To classify a single image, the neural network that predicts the corresponding severity score is applied once, and the resulting severity score is thresholded to determine the class label. Because the severity score is predicted by sigmoid activation, its range is in 0 to 1. Thus, we threshold the severity score at 0.5 to perform each binary classification task.

Statistical Analysis

Descriptive statistics, Welch's t test, and 2-way repeated-measures analyses of variance (ANOVAs) were performed with Microsoft Excel (Microsoft Corporation). Significance was set at $\alpha = 0.05$ for all tests. All values are presented as mean ± standard error of the mean. Where applicable, statistically significant differences between values are indicated on figures with asterisks.

Results

Experiment A

A neural network was trained with either comparison or class labels corresponding to 8, 16, and 24 images from the i-ROP dataset (Fig 4A, B). In both the normal versus abnormal and plus versus nonplus classification tasks, no statistically significant difference was calculated between models trained on either label type. Separately, a neural network was trained with either comparison or class labels corresponding to 3 and 6 images from the ICROP dataset (Fig 4C, D). In the normal versus abnormal task, the average AUC from training with comparison labels associated with 3 images was significantly higher than from training with class labels associated with the same number of images ($P = 0.008$, Welch's t test). For both classification tasks, training on comparison labels yielded significantly higher AUCs than training on class labels (2-way ANOVA: normal vs. abnormal, $F = 30.41$; main effect, $P = 0.0006$; plus vs. nonplus, $F = 5.83$; main effect, $P = 0.04$).

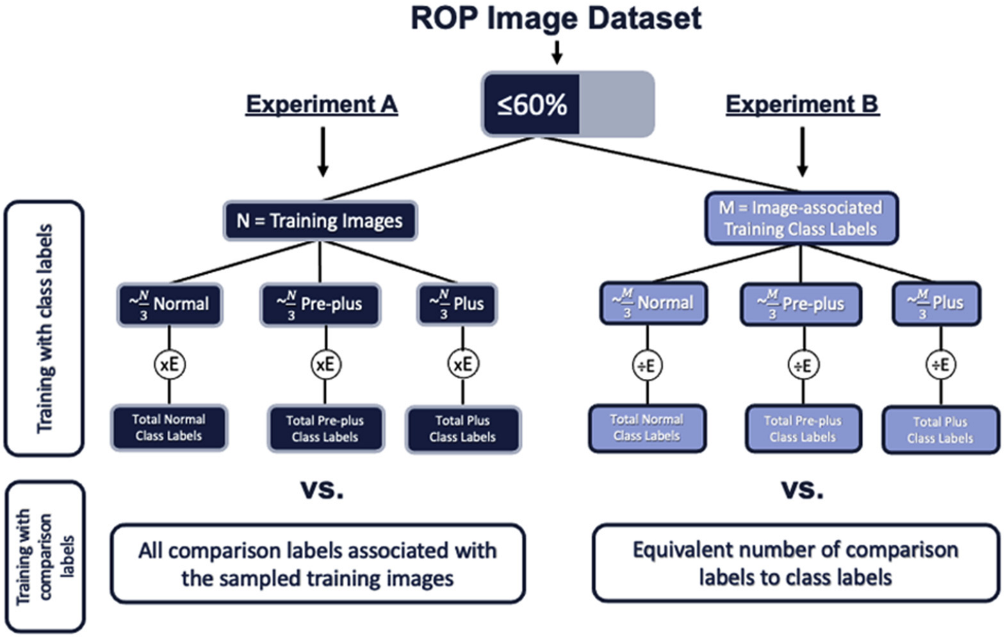
ROP Image Dataset
Experiment A
≤60%
Experiment B
N = Training Images
M = Image-associated Training Class Labels
Training with class labels
~N/3 Normal
~N/3 Pre-plus
~N/3 Plus
~M/3 Normal
~M/3 Pre-plus
~M/3 Plus
xE
÷E
Total Normal Class Labels
Total Pre-plus Class Labels
Total Plus Class Labels
vs.
Training with comparison labels
All comparison labels associated with the sampled training images
Equivalent number of comparison labels to class labels

Figure 3. Flow diagram showing a simplified depiction of neural network training between class and comparison labels in experiments A and B. Sixty percent of images from either the Imaging and Informatics in ROP (i-ROP) or International Classification of Retinopathy of Prematurity (ICROP) datasets were selected randomly. This selection then was balanced so as to achieve a near-even distribution of images represented by each of the 3 severity classes. In experiment A, the total number of class labels assigned to these images by expert graders then was used to train a neural network. Similarly, all comparison labels associated with the same images in this balanced training set were used to train a neural network for performance comparison. In experiment B, a set of class labels each corresponding to a single image in the balanced test set was used for training a neural network and was compared with a neural network trained on an equivalent number of comparison labels. E = total number of expert graders. ROP = retinopathy of prematurity.

Experiment B

A neural network was trained with 78, 156, 234, and 312 comparison or class labels from the i-ROP dataset (Fig 5A, B). In both classification tasks, the average AUC from training with 156 comparison labels was significantly higher than that measured from training with class labels (Welch's t test: normal vs. abnormal, $P = 0.002$; plus vs. nonplus, $P = 0.02$). Additionally, training on comparison labels in both classification tasks yielded significantly higher AUCs than training on class labels (2-way ANOVA: normal vs. abnormal, $F = 12.16$; main effect, $P = 0.003$; plus vs. nonplus, $F = 8.77$; main effect, $P = 0.009$). Separately, a neural network was trained with 70, 140, and 204 comparison or class labels from the ICROP dataset. In the normal versus abnormal task, the average AUC from training with 204 comparison labels was significantly higher than that measured from training with class labels ($P = 0.002$, Welch's t test; Fig 5C, D). Training on comparison labels yielded significantly higher AUCs than training on class labels in both classification tasks (2-way ANOVA: normal vs. abnormal: $F = 13.93$; main effect, $P = 0.003$; plus vs. nonplus, $F = 7.14$; main effect, $P = 0.02$).

Discussion

This study evaluated the relative performance of neural networks trained on either class or comparison labels for classification of disease severity in ROP fundus images. Given the same number of represented images, as in experiment A, learning from comparison labels generated more accurate neural networks, achieving statistical significance in both classification tasks for training sets derived from the ICROP dataset. This observation may be explained in part by the fact that pairwise comparisons allow for multiple comparison labels to be associated with a single image, as opposed to a single diagnostic class label. With more labels available for training per image, a neural network therefore may have a deeper pool of samples from which it may be trained and validated. The use of comparison labels in this way offers a potential solution for training image classification models with small datasets.

Because the number of labels available for training per image was greater using comparisons in experiment A, we additionally investigated whether network performance may differ when training on equal numbers of label type. Given the same number of labels, as in experiment B, neural networks using comparison labels achieved higher AUC, exhibiting statistically significant main effects in both classification tasks with both datasets. This may be explained by prior observations that comparison labels elicit less intergrader variability, or noise, compared with that of class labels.[10,20,21]

Although 2-way ANOVAs were useful in projecting a main effect of treatment, or training label type, statistically significant differences per Welch's t test were not calculated consistently between models trained on the same number of images or labels. However, significant differences were

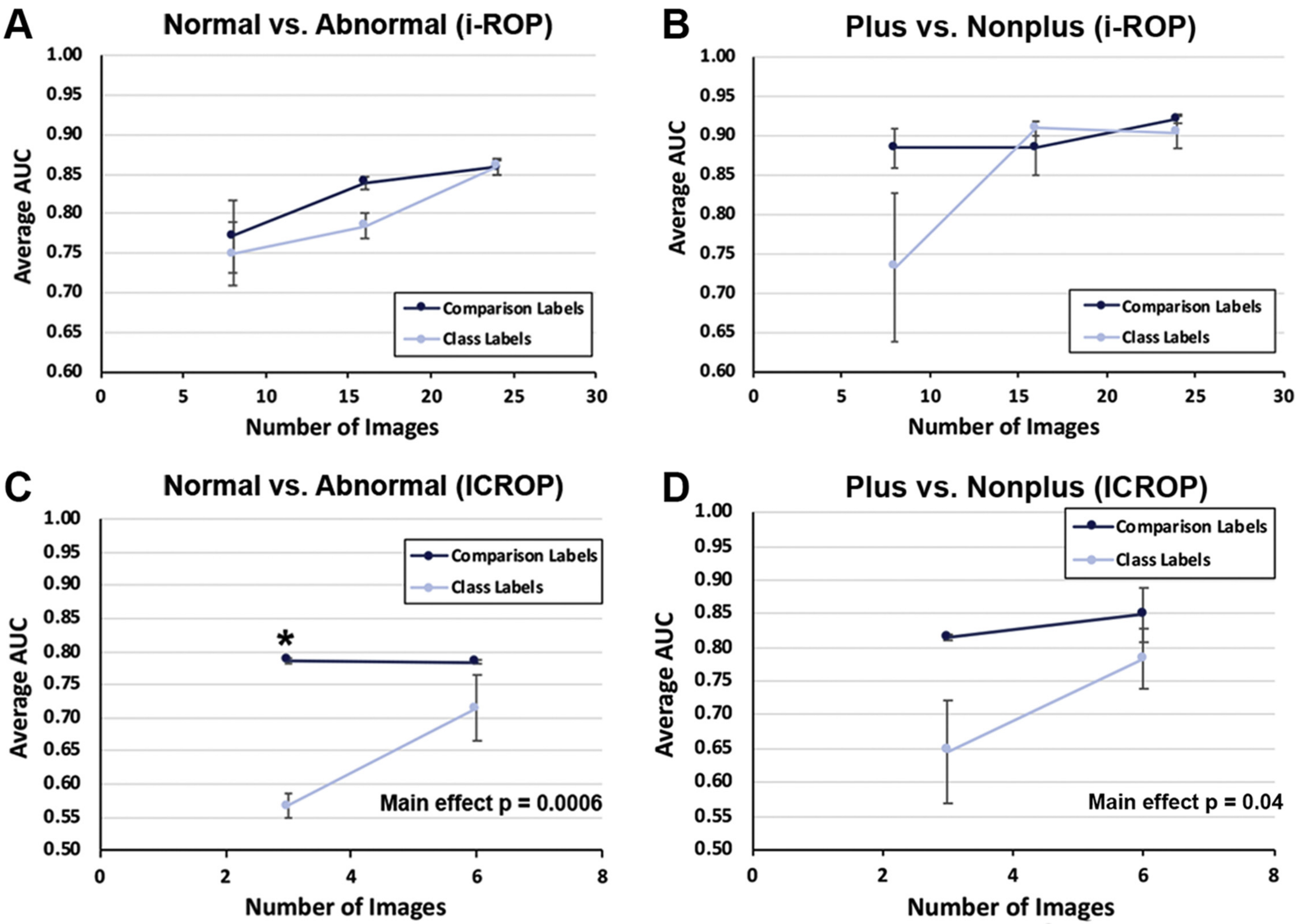

Figure 4. Line graphs showing experiment A neural network performance. **A**, **B**, Normal versus abnormal (**A**) and plus versus nonplus (**B**) classification tasks from models trained on class or comparison labels corresponding to images within the Imaging and Informatics in ROP (i-ROP) dataset. No statistically significant difference was calculated between models trained on either label type. **C**, **D**, Classification performances from models trained on class or comparison labels corresponding to images within the Classification of Retinopathy of Prematurity (ICROP) dataset. Training on comparison labels yielded significantly higher area under the receiver operating characteristic curves (AUCs) than training on class labels (2-way analysis of variance: normal vs. abnormal: $F = 30.41$; main effect, $P = 0.0006$; plus vs. nonplus: $F = 5.83$; main effect, $P = 0.04$). In the normal versus abnormal task (**C**), the average AUC from training with comparison labels associated with 3 images was significantly higher than from training with class labels associated with the same number of images ($P = 0.008$, Welch's t test).

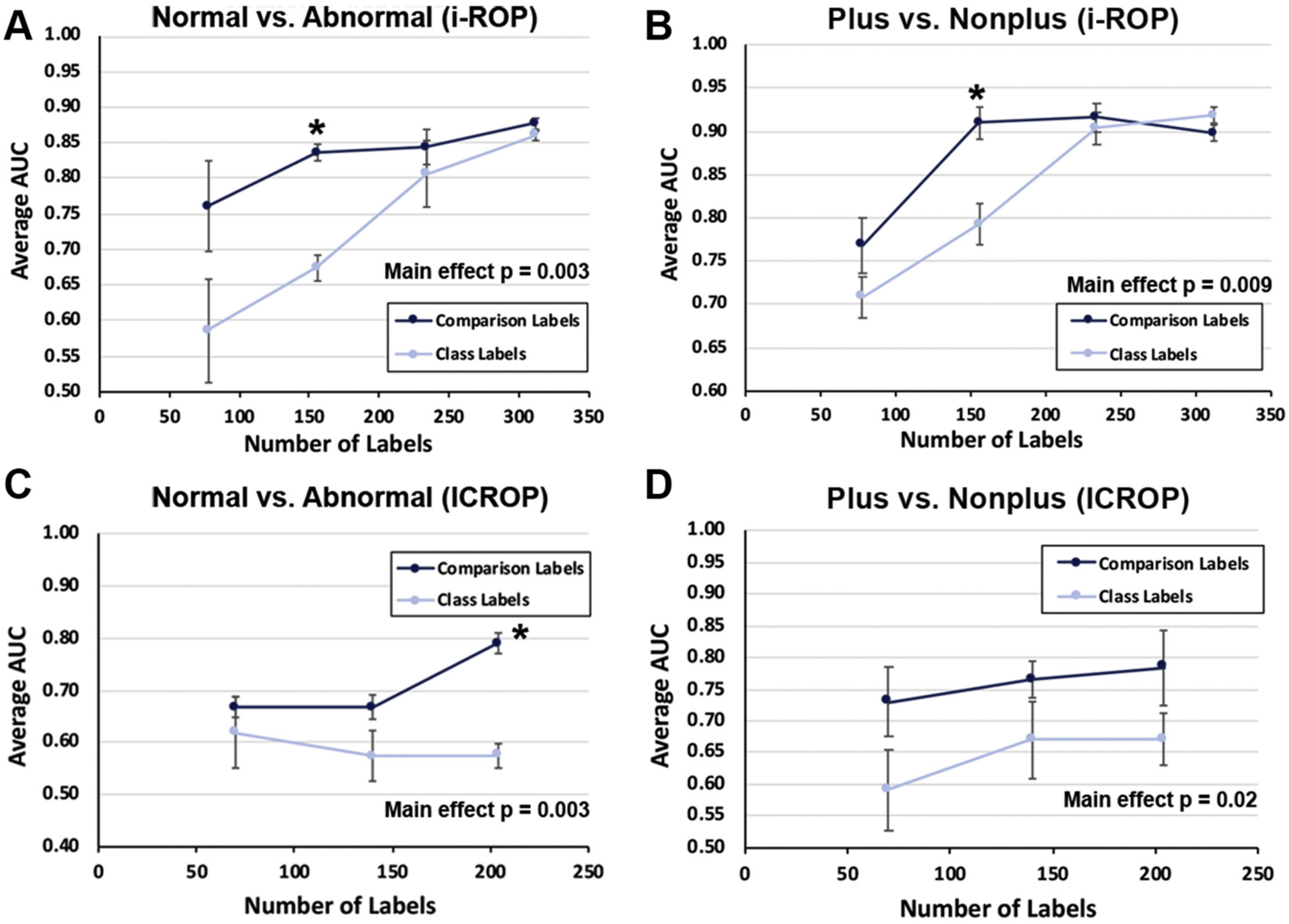

Figure 5. Line graphs showing experiment B neural network performance. **A**, **B**, Normal versus abnormal (**A**) and plus versus nonplus (**B**) classification tasks from models trained on class or comparison labels within the Imaging and Informatics in ROP (i-ROP) dataset. **A**, **B**, Average area under the receiver operating characteristic curve (AUC) from training with 156 comparison labels was significantly higher than that measured from training with class labels (Welch's *t* test: normal vs. abnormal, $P = 0.002$; plus vs. nonplus, $P = 0.02$). Training on comparison labels yielded significantly higher AUCs than training on class labels (2-way analysis of variance [ANOVA]: normal vs. abnormal: $F = 12.16$; main effect, $P = 0.003$; plus vs. nonplus: $F = 8.77$; main effect, $P = 0.009$). **C**, **D**, Classification performances from models trained on class or comparison labels corresponding to images within the International Classification of Retinopathy of Prematurity (ICROP) dataset. Training on comparison labels yielded significantly higher AUCs than training on class labels in both classification tasks (normal vs. abnormal: 2-way ANOVA: $F = 13.93$; main effect, $P = 0.003$; plus vs. nonplus: $F = 7.14$; main effect, $P = 0.02$). In the normal versus abnormal task (**C**), the average AUC from training with 204 comparison labels was significantly higher than that measured from training with class labels ($P = 0.002$, Welch's *t* test).

observed most frequently between models trained with fewer samples. This not only supports the presumed greater efficiency of training with comparison labels, but also a diminishing difference in performance between networks trained on either label type as the size of the training set increases. At the greatest sizes of training set used, learning by both label types achieved AUC values comparable with those achieved by other deep learning models applied for medical imaging classification.[22,23] However, networks trained on comparison labels approached these performance levels earlier, as the size of training set became incrementally greater. The usefulness of this approach when high-quality models are required in the setting of limited data therefore may be circumstantial, because the performance of networks trained on either label type may achieve similar levels of performance when trained in data-replete settings.

Although the use of comparison labels offers an alternative solution to training with noisy, small datasets, they are more labor intensive, or expensive, to obtain. Whereas a single diagnostic class label may be assigned per image per grader, comparison labels require graders to label all pairwise comparisons of the images in the set independently. To facilitate this process, we used an internally developed image severity assessment platform that presented graders with 2 images to compare and incorporated responses into an Elo algorithm to generate rankings.

Ordinal classifications long have been used in medicine to indicate severity in continuous disease processes. In the context of ROP, International Classification of ROP (ICROP) criteria are used conventionally to derive subclassifications of zone (I—III), stage (0—5), and plus disease status (present or not) from subjective and qualitative assessment of disease features that direct treatment and guide clinical trials.[24] However, recognition of ROP classifications as checkpoints on a disease continuum is increasing, most recently exemplified in 2021 by the update on ICROP, third edition, which formally recognizes preplus and plus disease as part of a continuous spectrum of disease.[25] As the frameworks for ROP disease classification increasingly reflect its underlying mechanisms, so must the appropriate models be applied to the classification tasks at hand. The use of comparison labels in training may be a more fitting way to train image classifiers tasked with assigning ordinal terms that individually represent a range of severity on the ROP spectrum. As neural networks are implemented for classification in other continuous disease models, this approach to training should be considered.

Limitations

The findings and interpretation of this study are limited by the time and computing power required both to acquire more expert labels per image and to perform multiple repetitions of experiments. Access to datasets of labels that are both more numerous per image and distributed more evenly between severity class per grader would permit a wider range of training set sizes after the balancing process and would characterize AUC curve profiles more accurately. Our analysis of the ICROP dataset in experiment A was limited in this way, with a maximum training set size of 6 images and only 2 average AUCs because of the minimal difference between possible iterations of training set size. Furthermore, the ability to perform more experiment repetitions presumably would reduce variability between multiple trials of testing at a given set size and would draw subjectively observable differences between groups toward statistical significance. This also may enable more informed choices of methods for statistical analysis. Performing only 3 repetitions of each experiment per size of training set precluded the assessment of normality in our data. Although more fitting methods for comparison of a continuous, skewed variable interest between 2 independent samples were considered, such tests as the Mann—Whitney *U* test generally require larger sample sizes.[26] The Welch's *t* test for comparison of 2 independent samples of unequal variance thus was chosen, conceding the assumption of a normal distribution. We additionally used 2-way repeated-measures ANOVAs to estimate a main effect of treatment (i.e., training label type) on the dependent variable AUC across the independent variable of training set size. To justify this approach, we again had to permit the assumption of normality, as well as repeated measures design. In this case, we interpreted repeated measures to involve the multiple measurements of the dependent variable AUC taken on the same subjects (i.e., neural network architectures) under different conditions (i.e., training set size).[27]

In conclusion, the potential of neural networks to generate predictions in the context of medical image classification often is limited by datasets of modest size and quality. We propose an alternative approach to training models with labels generated from pairwise comparisons of disease severity between images within the dataset. Our data indicate that grading images by comparison generates labels that are more abundant and informative per image than diagnostic class labels. This method may offer a solution for improving the efficiency of training models and training highly accurate models in data-scarce settings.

Acknowledgments

The authors thank the following ICROP members who additionally participated as expert graders: Alistair Fielder, Anand Vinekar, Andrea Zin, Andreas Stahl, Andrés Kychenthal, Anna Ells, Antonio Capone, Jr., Audina Berrocal, Birgit Lorenz, Brian Fleck, Cynthia Toth, David Wallace, Domenico Lepore, Dupe Ademola-Popoola, Gerd Holmstrom, Gil Binenbaum, Graham Quinn, James Reynolds, Linda Visser, Maria Ana Martinez-Castellanos, Mary Elizabeth Hartnett, Michael Blair, Michael Shapiro, Parag Shah, Şengül Özdek, Shuan Dai, Shunji Kusaka, Wei-Chi Wu, William Good, and Yi Chen.

Footnotes and Disclosures

Originally received: August 26, 2021.
Final revision: December 13, 2021.
Accepted: January 26, 2022.
Available online: February 2, 2022. Manuscript no. XOPS-D-21-00157.

[1] Department of Ophthalmology, Oregon Health & Science University, Portland, Oregon.

[2] Department of Electrical and Computer Engineering, Northeastern University, Boston, Massachusetts.

[3] Department of Radiology, Athinoula A. Martinos Center for Biomedical Imaging Clinical Computational Neuroimaging Group, Charlestown, Massachusetts.

[4] Department of Ophthalmology, University of Illinois at Chicago College of Medicine, Chicago, Illinois.

[5] National Eye Institute, National Institutes of Health, Bethesda, Maryland.

Presented at: Association for Research in Vision and Ophthalmology Annual Meeting, Virtual Meeting, May 2021.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The author(s) have made the following disclosure(s):

R.V.P.C.: Consultant − Novartis; Financial support − Genentech, Regeneron; Scientific board − Phoenix Technology; Equity owner − Siloam Vision.

M.F.C.: Consultant − Novartis; Financial support − Genentech; Equity owner − InTeleretina, LLC

J.P.C.: Consultant − Boston AI; Financial support − Genentech; Equity owner − Siloam Vision.

Supported by the National Institutes of Health, Bethesda, Maryland (grant nos.: R01EY19474, R01 EY031331, R21 EY031883, and P30 EY10572); Genentech, Inc., San Francisco, CA (investigator-initiated grant); and Research to Prevent Blindness, Inc., New York, New York (unrestricted departmental funding and Career Development Award [J.P.C.]). The sponsors or funding organizations had no role in the design or conduct of this research.

HUMAN SUBJECTS: Human subjects were included in this study. This study was approved by the Institutional Review Board at Oregon Health & Science University and all participating institutions (Beaumont Health, Cedars Sinai Medical Center, Children's Hospital of Los Angeles, Columbia University Medical Center, Weill Cornell Medical Center, University of Miami Health System) in the i-ROP cohort study. All research adhered to the tenets of the Declaration of Helsinki. All parents of infants whose images were included in the datasets provided informed consent.

No animal subjects were included in this study.

Author Contributions:

Conception and design: Hanif, Yıldız, Tian, Erdoğmuş, Ioannidis, Dy, Kalpathy-Cramer, Jonas, Chiang, Campbell

Analysis and interpretation: Hanif, Yıldız, Tian

Data collection: Yıldız, Tian, Kalkanlı, Ostmo, Chan

Obtained funding: N/A

Overall responsibility: Hanif

Abbreviations and Acronyms:

ANOVA = analysis of variance; **AUC** = area under the receiver operating characteristic curve; **ICROP** = International Classification of Retinopathy of Prematurity; **i-ROP** = Imaging and Informatics in ROP; **ROP** = retinopathy of prematurity.

Keywords:

Artificial intelligence, Deep learning, Labels, Neural networks, Retinopathy of prematurity.

Correspondence:

J. Peter Campbell, MD, MPH, Department of Ophthalmology, Oregon Health & Science University, 515 SW Campus Drive, Portland, OR 97239. E-mail: campbelp@ohsu.edu.

References

1. Choi RY, Coyner AS, Kalpathy-Cramer J, et al. Introduction to machine learning, neural networks, and deep learning. *Transl Vis Sci Technol*. 2020;9(2):14.
2. Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15(1):1929−1958.
3. Schmidhuber J. Deep learning in neural networks: an overview. *Neural Networks*. 2015;61:85−117.
4. Chang K, Beers AL, Brink L, et al. Multi-institutional assessment and crowdsourcing evaluation of deep learning for automated classification of breast density. *JACR*. 2020;17(12):1653−1662.
5. Dunnmon JA, Yi D, Langlotz CP, et al. Assessment of convolutional neural networks for automated classification of chest radiographs. *Radiology*. 2018;290(2):537−544.
6. Yıldız İ, Tian P, Dy J, et al. Classification and comparison via neural networks. *Neural Networks*. 2019;118:65−80.
7. Sun W-T, Chao T-H, Kuo Y-H, Hsu W. Photo filter recommendation by category-aware aesthetic learning. *IEEE Transactions on Multimedia*. 2016;19(8): 1870−1880.
8. Campbell JP, Kalpathy-Cramer J, Erdogmus D, et al. Plus disease in retinopathy of prematurity: a continuous spectrum of vascular abnormality as a basis of diagnostic variability. *Ophthalmology*. 2016;123(11):2338−2344.
9. Chiang MF, Jiang L, Gelman R, et al. Interexpert agreement of plus disease diagnosis in retinopathy of prematurity. *Arch Ophthalmol*. 2007;125(7):875−880.
10. 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(11):2345−2351.
11. Bradley RA, Terry ME. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*. 1952;39(3/4):324−345.
12. Cattelan M. Models for paired comparison data: a review with emphasis on dependent data. *Stat Sci*. 2012;27(3):412−433.
13. Bromley J, Guyon I, LeCun Y, et al. Signature verification using a "Siamese" time delay neural network. In: Cowan JD, Tesauro G, Alspector J, eds. *Proceedings of the 6th International Conference on Neural Information Processing Systems*. Denver, CO: Morgan Kaufmann Publishers Inc.; 1993:737−744.
14. Ryan MC, Ostmo S, Jonas K, et al. Development and evaluation of reference standards for image-based telemedicine diagnosis and clinical research studies in ophthalmology. *AMIA Annu Symp Proc*. 2014;2014:1902−1910.
15. Wallace DK, Freedman SF, Hartnett ME, Quinn GE. Predictive value of pre-plus disease in retinopathy of prematurity. *Arch Ophthalmol*. 2011;129(5):591−596.
16. 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(7): 803−810.
17. Falk T, Mai D, Bensch R, et al. U-Net: deep learning for cell counting, detection, and morphometry. *Nat Methods*. 2019;16(1):67−70.

18. Szegedy C, Liu W, Jia Y, et al., eds. *Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston, MA: Institute of Electrical and Electronics Engineers (IEEE); 2015:1−9. https://doi.org/10.1109/CVPR.2015.7298594.
19. Deng J, Dong W, Socher R, et al., eds. *ImageNet: a large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: Institute of Electrical and Electronics Engineers (IEEE); 2009:248−255. https://doi.org/10.1109/CVPR.2009.5206848.
20. Stewart N, Brown GD, Chater N. Absolute identification by relative judgment. *Psychol Rev*. 2005;112(4):881−911.
21. Brun A, Hamad A, Buffet O, Boyer A, eds. *Towards preference relations in recommender systems. Workshop on Preference Learning (PL2010) in European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases—ECML-PKDD*. Barcelona, Spain: Eyke Hüllermeier and Johannes Fürnkranz; 2010.
22. Chan S, Reddy V, Myers B, et al. Machine learning in dermatology: current applications, opportunities, and limitations. *Dermatol Ther (Heidelb)*. 2020;10(3):365−386.
23. Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol*. 2019;103(2):167.
24. The Committee for the Classification of Retinopathy of Prematurity. An international classification of retinopathy of prematurity. *Arch Ophthalmol*. 1984;102(8):1130−1134.
25. Chiang MF, Quinn GE, Fielder AR, et al. International Classification of Retinopathy of Prematurity, Third Edition. *Ophthalmology*. 2021;128(10):e51−e68.
26. Nachar N. The Mann-Whitney U: a test for assessing whether two independent samples come from the same distribution. *Tutor Quant Methods Psychol*. 2008;4:13−20.
27. Salkind NJ. *Encyclopedia of Research Design*. 1st ed. Newbury Park, CA: SAGE; 2010.