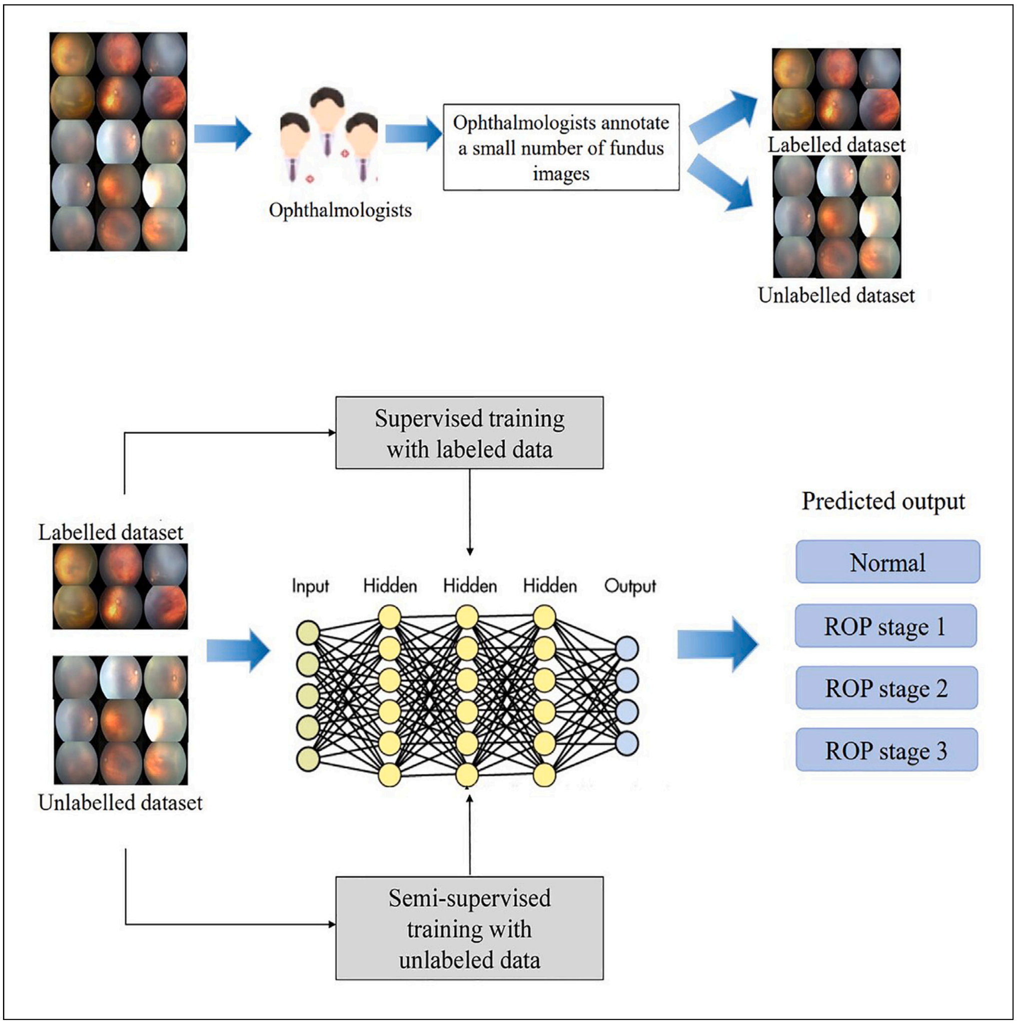
iScience

CellPress
OPEN ACCESS

Article

Development and validation of a semi-supervised deep learning model for automatic retinopathy of prematurity staging



Wei Feng, Qiujing Huang, Tong Ma, Lie Ju, Zongyuan Ge, Yuzhong Chen, Peiquan Zhao

hqj1010@126.com (Q.H.)
zhaopeiquan@xinhuamed.com.cn (P.Z.)

Highlights

A new semi-supervised classification model to utilize unlabeled data for ROP staging

Two consistency losses to efficiently mine information from unlabeled data

Experiments verify that our method improves classification performance

Feng et al., iScience *27*, 108516
January 19, 2024 © 2023 The Authors.
https://doi.org/10.1016/j.isci.2023.108516

iScience

CellPress
OPEN ACCESS

Article

Development and validation of a semi-supervised deep learning model for automatic retinopathy of prematurity staging

Wei Feng,[1,2,5] Qiujing Huang,[3,4,5,*] Tong Ma,[1] Lie Ju,[1,2] Zongyuan Ge,[2] Yuzhong Chen,[1] and Peiquan Zhao[3,6,*]

SUMMARY

Retinopathy of prematurity (ROP) is currently one of the leading causes of infant blindness worldwide. Recently significant progress has been made in deep learning-based computer-aided diagnostic methods. However, deep learning often requires a large amount of annotated data for model optimization, but this requires long hours of effort by experienced doctors in clinical scenarios. In contrast, a large number of unlabeled images are relatively easy to obtain. In this paper, we propose a new semi-supervised learning framework to reduce annotation costs for automatic ROP staging. We design two consistency regularization strategies, prediction consistency loss and semantic structure consistency loss, which can help the model mine useful discriminative information from unlabeled data, thus improving the generalization performance of the classification model. Extensive experiments on a real clinical dataset show that the proposed method promises to greatly reduce the labeling requirements in clinical scenarios while achieving good classification performance.

INTRODUCTION

Retinopathy of prematurity (ROP) is a proliferative disorder of the developing retina in premature infants.[1] It continues to be a major cause of childhood blindness worldwide. The RetCam wide-field digital retinal imaging system (Natus Medical Incorporated, San Carlos, CA, USA) is currently one of the primary instruments utilized for ROP screening. It has a 130° wide-field lens. After dilation of the pupils, it enables for multidirectional imaging of the retina. Furthermore, the RetCam system provides image storage and transmission in addition to fundus photography, setting the groundwork for telemedicine and artificial intelligence applications. Ophthalmologic examination of preterm infants necessitates frequent and close monitoring, resulting in a massive workload of manual image reading.[2] Clinically, however, there is subjectivity and diagnostic variation in the diagnosis of ROP stage 1, 2, and 3.[3,4] Failure to detect and treat ROP promptly results in late-stage ROP, which can cause low vision and even blindness. Furthermore, experienced pediatric ophthalmologists are in short supply, with the majority concentrated in large cities or large medical centers. Infants with ROP who live in remote areas must travel long distances for referrals, which delays treatment. Moreover, they may be ineligible for referral in some cases due to poor general health. As a result, telemedicine and computer-assisted ROP fundus images reading are extremely useful.

In recent years, deep learning techniques have made breakthroughs in various fields such as computer vision,[5,6] natural language processing,[7,8] and speech recognition.[9,10] In the field of ROP image analysis, many deep learning models have also been proposed for computer-aided screening and diagnosis. Peng et al.[11] proposed a new deep learning model for ROP staging that utilizes parallel feature extraction, deep feature fusion, and sequential classifiers to extract a richer feature representation. Wang et al.[12] proposed a two-stage deep learning model, Id-Net and Gr-Net, for the ROP recognition and ROP staging tasks, respectively. Lei et al.[13] introduced channel attention and spatial attention mechanisms to improve the performance of ROP detection. However, deep learning methods often require large amounts of annotated data for model training. In clinical scenarios, annotating large amounts of data is often time-consuming and laborious, even for experienced doctors. For the annotation of ROP fundus images, the clinician usually determines the stage of ROP based on the shape and size of the ridge in the ROP fundus images. In this way, a stage 4 or 5 ROP fundus image is easily distinguished, whereas a normal fundus image, a stage 1 ROP image, or a stage 2 ROP image may be misinterpreted by the clinician because it is in the early stages of the disease and the disease is not very distinct, thus increasing the burden on the clinician. Instead, we often have access to large amounts of unlabeled data, which are relatively easy to obtain and may be beneficial to the classification performance of the model.

[1]Beijing Airdoc Technology Co., Ltd, Beijing 100089, China
[2]Faculty of Engineering, Monash University, Melbourne, VIC 3000, Australia
[3]Department of Ophthalmology, Xinhua Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 200092, China
[4]Department of Ophthalmology, Rainbow Children's Clinic, Shanghai 200010, China
[5]These authors contributed equally
[6]Lead contact
*Correspondence: hqj1010@126.com (Q.H.), zhaopeiquan@xinhuamed.com.cn (P.Z.)
https://doi.org/10.1016/j.isci.2023.108516

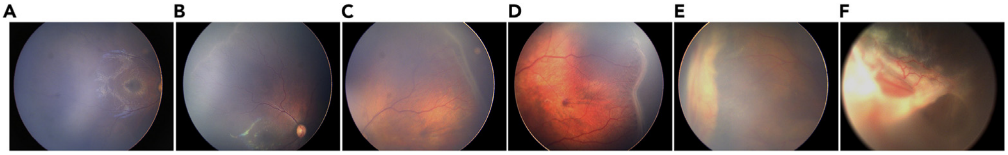


Figure 1. Normal fundus images and ROP fundus images of different stages

(A) Normal fundus.

(B) ROP stage1.

(C) ROP stage2.

(D) ROP stage3.

(E) ROP stage4.

(F) ROP stage5.

Semi-supervised learning is proposed to minimize the need for annotation as data annotation is time-consuming and laborious. Given a small amount of labeled data, semi-supervised learning improves the classification performance of the model by utilizing a large amount of unlabeled data.[14,15] The current state-of-the-art semi-supervised learning strategy is consistency regularization, which does not rely on data labels and improves the classification performance of the model by constraining the consistency of model predictions under different perturbations of the data.[16–18] There has been some research work on the use of semi-supervised learning for medical image analysis tasks.[19–21] For example, Zhao et al.[22] proposed a two-stage cascade network for cup-to-disc ratio estimation based on semi-supervised learning. Adal et al.[23] proposed a semi-supervised learning method using a small number of manual markers and a large number of unlabeled fundus images for the automatic detection of microaneurysms (MAs). Liu et al.[24] proposed a semi-supervised conditional generative adversarial nets (GANs) for the joint segmentation of optic cup and disc. Liu et al.[25] proposed a semi-supervised deep learning model using relationship information among different samples to improve the classification performance of the model and achieved promising performance on skin disease classification and thorax disease classification. However, there are still few studies on semi-supervised methods for ROP fundus images classification, and, in addition, these methods do not take into account the semantic structure correlation between different ROP stages. In this paper, we propose a semi-supervised deep learning method for ROP staging that can make use of a small amount of labeled data and a large amount of unlabeled data to alleviate the annotation cost of doctors and improve the classification performance of the model. Specifically, firstly, inspired by the consistency regularization technique in semi-supervised learning, we argue that the predictions of the model should not be affected by the small perturbations imposed on the data. We therefore propose a prediction consistency loss, which forces the model to produce consistent prediction outputs for the original data under different perturbations, thus improving the generalization performance of the model. In addition, considering that there is some correlation between different stages of ROP, for example, as shown in Figure 1, stage 1 ROP tends to manifest as a white demarcation line between vascular and non-vascular regions in the posterior pole of the retina. Stage 2 shows a further widening and elevation of the demarcation line and a crestal bulge. In stage 3, the crestal bulge becomes more pronounced and is accompanied by neovascularization. Stages 4 and 5 often show a further retinal detachment, necessitating ocular ultrasonography. Some semantic structure correlation between the different stages of ROP can be found, which may be useful for the classification performance of the model. We therefore propose a semantic structure consistency loss that forces the semantic structure relationships to remain consistent under different perturbations, thus further extracting rich semantic information from the unlabeled data to enhance the classification performance.

RESULTS

Construction of the semi-supervised model for ROP image classification

Figure 2 depicts our proposed semi-supervised deep learning classification model for ROP fundus images classification. The model structure is similar to the popular semi-supervised classification model mean teacher,[15] containing a student model and a teacher model, which are identical in structure. The parameters of the teacher model are optimized by exponential moving average algorithm on the parameters of the student model.[15] The parameters of the student model are optimized by supervised cross-entropy loss on labeled data and consistency loss on unlabeled data. The consistency loss is calculated from the output of the student model and the output of the teacher model. It consists of prediction consistency loss and semantic structure consistency loss. The prediction consistency loss enhances the generalization performance of the model by encouraging the two models to maintain the same prediction output for the same input image with different data augmentation to mine the information in the unlabeled image. In addition, there is some semantic relevance between different stages of the ROP, and this semantic relevance may help improve model classification performance. We use semantic structure consistency loss to further encourage the model to maintain the same semantic structure for the fundus images after data augmentation to further mine useful information from the unlabeled images to help improve the classification performance.

Calculation of consistency loss

The prediction consistency loss can be formulated as

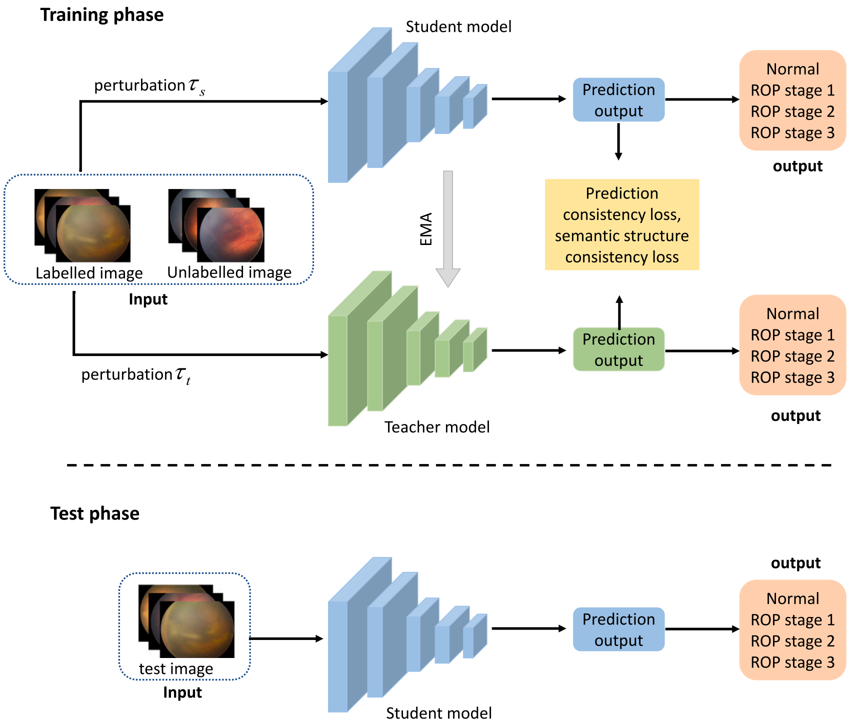


Figure 2. Illustration of the framework of the proposed semi-supervised deep learning classification model

Labeled and unlabeled fundus images are input to the model. The student model was optimized based on supervised loss on labeled data and prediction consistency loss and semantic structure consistency loss on all data. The teacher model was optimized based on an exponential moving average (EMA) algorithm. We used the student model for inference in the testing phase.

$$L_c = \frac{1}{N_l+N_u}\sum_{i=1}^{N_l+N_u}(E(x_i) < \mu)\|f_s(x_i,\theta_s,\tau_s) - f_t(x_i,\theta_t,\tau_t)\|_2^2 \tag{Equation 1}$$

where τ_s and τ_t represent the random perturbations imposed on the original fundus images, f_s and f_t denote the student and teacher models, and θ_s and θ_t denote the weights of the student and teacher models, respectively. N_l and N_u are the numbers of labeled images and unlabeled images, respectively. $E(x_i)$ represents the entropy of the teacher model's prediction for sample x_i , and μ is the threshold. Here, since the model is less stable during the early training period, in order to allow the model to gradually learn from meaningful and reliable targets, we consider only those samples that are reliable. We compute the information entropy of the teacher model's predictions for each sample and filter out those with high entropy (uncertain, noisy samples).

For the calculation of semantic structure consistency loss, we first feed a batch of fundus images into the classification model and obtain its deep feature representation $S \in \mathbb{R}^{B\times D\times H\times W}$, where B represents the batch size. D is the number of feature channels. H and W represents the spatial dimension of the deep feature.

We then reshape the deep feature to $Q \in \mathbb{R}^{B\times HWD}$. We compute the case-wise gram matrix[26] between the samples in the batch $G = Q \cdot Q^T$, where $G_{i,j}$ represents the similarity between sample i and sample j. Intuitively, it can reflect the semantic structure correlation between samples in the current batch. We then normalize each row of the semantic structure correlation matrix G: $P = \left[\frac{Q_1}{\|Q_1\|_2}, \ldots, \frac{Q_B}{\|Q_B\|_2}\right]^T$.

We calculate the semantic structure correlation matrix for the student model and the teacher model separately, and the semantic structure consistency loss can be expressed as

$$L_s = \frac{1}{B}\sum_{i=1}^{B}\left\|P^s(x_i,\theta_s,\tau_s) - P^t(x_i,\theta_t,\tau_t)\right\|_2^2 \tag{Equation 2}$$

where P^s and P^t represent the semantic structure correlation matrices of the student and teacher models, respectively.

Finally, we can obtain the overall training loss of our proposed semi-supervised classification model:

$$L_{total} = L_{ce} + \lambda(L_c + L_s) \tag{Equation 3}$$

where L_{ce} represents the classification cross-entropy loss computed on a small amount of labeled data, and λ is balance coefficient controlling the weights of the two consistency losses.

Table 1. Characteristics of the training and test dataset

No. of images	Training dataset	Test dataset
Normal	3,171	92
Stage 1	744	67
Stage 2	426	154
Stage 3	455	73
Total	4,796	386

Dataset description

We used fundus images from 473 infants at Xinhua Hospital as the training set, with a sex ratio of 1.2:1 (male:female). The mean gestational age was 29.86 ± 2.47 weeks, and the mean birth weight was 1379.95 ± 412.95 grams. The dataset contained 4,796 fundus images, of which 3,171 were normal, 744 were stage 1 ROP, 426 were stage 2 ROP, and 455 were stage 3 ROP. To evaluate the performance of the model, we used fundus images of 62 infants, also collected from Xinhua Hospital, as the test set, with a sex ratio of 1.4:1 (male:female). The mean gestational age was 29.29 ± 2.94 weeks, and the mean birth weight was 1381.77 ± 451.75 grams. The dataset contained 386 fundus images, of which 92 were normal, 67 were stage 1 ROP, 154 were stage 2 ROP, and 73 were stage 3 ROP. The image data were collected and analyzed with the approval of the Xinhua Hospital Ethics Committee and in compliance with the tenets of the Declaration of Helsinki. Detailed statistics of the dataset are shown in Table 1.

Experimental setup and implementation details

We resized all fundus images into 224 × 224 as input to the classification model. We used ResNet50 as our network backbone and initialized it with the pre-trained weights from the ImageNet dataset. The student model and the teacher model have the same structure. We use area under the receiver operating characteristic curve (AUC), Accuracy, Sensitivity, and Specificity as evaluation metrics to evaluate the performance of different algorithms. We used Python and PyTorch to conduct the experiments. All experiments were run on two NVIDIA 3090 Ti GPUs. We trained a total of 200 epochs using the Stochastic Gradient Descent (SGD) optimizer, with a learning rate set to 1e-4 and a batch size of 128, containing 32 labeled samples and 96 unlabeled samples. For random perturbations we use random horizontal/vertical flipping, random rotation, and color jittering. In addition we add a dropout layer before the global pooling layer of the ResNet50 network and set the dropout rate to 0.2. The threshold μ is set to 0.9. The balancing coefficient λ is set to 1. We use a Gaussian warm-up function $\lambda_t = 1 * e^{(-5(1-t/T)^2)}$ to control the two consistency losses, which gradually increase from 0 to 1 in the first T epochs and then fixed. This approach avoids the model being dominated by the two consistency losses at the beginning of training, where the model's predictions are unreliable and therefore not conducive to consistent training.

ROP staging results

We first trained the model using 30% (1,438) of the labeled fundus images in the training set, with the remaining 70% (3,358) fundus images being unlabeled. Model performance with only 30% (1438) labeled fundus images (without 70% unlabeled images) as the training set is also presented as the lower bound (baseline_0.3) for comparison. As can be seen from Table 2, when only 30% of the training data are used for model training, the classification performance of the model is poor, with an average AUC of 0.8226, an average Accuracy of 0.7590, an average

Table 2. Classification performance of different algorithms on the test set

		Metrics			
Method	Class	AUC	Accuracy	Sensitivity	Specificity
Baseline_0.3	Normal	0.9918	0.9378	0.9783	0.9252
Ours		**0.9996**	**0.9715**	**1.0000**	**0.9626**
Baseline_0.3	ROP stage 1	0.7679	**0.7280**	0.5821	**0.7586**
Ours		**0.8402**	0.7202	**0.8507**	0.6928
Baseline_0.3	ROP stage 2	0.7185	0.6632	**0.6429**	0.6767
Ours		**0.7655**	**0.6839**	0.5844	**0.7500**
Baseline_0.3	ROP stage 3	0.8126	0.7073	**0.8219**	0.6805
Ours		**0.8470**	**0.7824**	0.7123	**0.7987**
Baseline_0.3	Average	0.8226	0.7590	0.7562	0.7602
Ours		**0.8630**	**0.7895**	**0.7868**	**0.8010**

Bold in the tables represent better performance results.

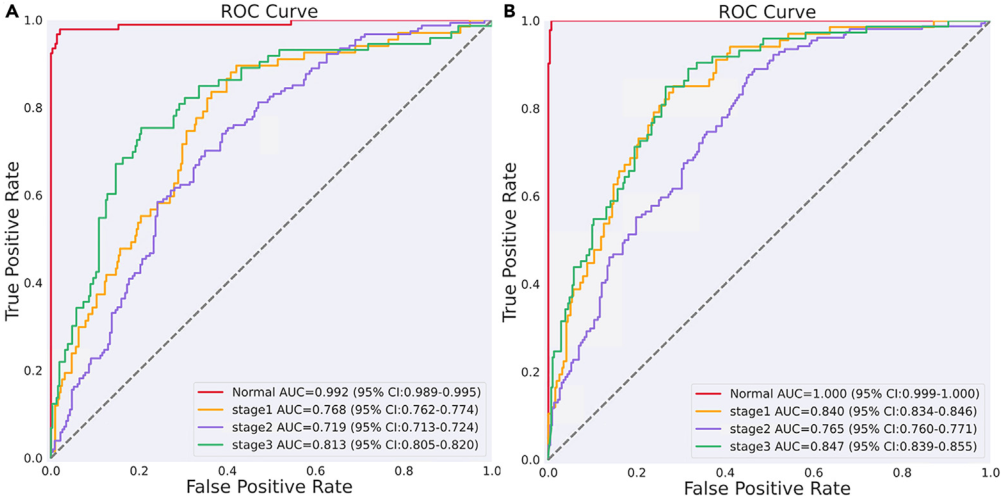


Figure 3. Receiver operating characteristic curves for different methods

(A) Baseline_0.3, (B) ours.

Sensitivity of 0.7562, and an average Specificity of 0.7602. Our approach forces the model to be consistent not only in its predictions for individual samples but also in its semantic structure between samples, thus achieving better classification performance with an average AUC improvement of 4.04%, an average Accuracy improvement of 3.05%, an average Sensitivity improvement of 3.06%, and an average Specificity improvement of 4.08%. As can be seen from the receiver operating characteristic curves in Figure 3 and the confusion matrix in Figure 4, our approach achieves performance gains in all categories, which further validates the superiority of our approach.

Effect of different proportions of labeled data

To further examine the classification performance of our proposed semi-supervised deep learning classification model with different proportions of labeled data, we re-run the experiment by varying the number of labeled samples. We also report the model performance under full supervision, i.e., 100% (4,796) of the labeled fundus images as the training set, as an upper bound (Oracle) for the comparison. Baseline_0.05, Baseline_0.1, Baseline_0.15, Baseline_0.2, Baseline_0.3, and Baseline_0.5 denote that only 5%, 10%, 15%, 20%, 30%, and 50% of the labeled data were used, respectively, and none of them used unlabeled data in the training process. As shown in Table 3, the performance of the model improves as the sample of labeled training data increases. Furthermore, it can be found that, when we use 50% of the labeled data, the performance of the proposed algorithm is already close to that of fully supervised (using 100% of the labeled data). This suggests that the proposed method can make full use of unlabeled data to improve the classification performance of the model and significantly reduce the labeling burden on the doctors.

Visualization of the evolution of the semantic structure correlation matrix

To further understand the behavior of the proposed semantic structure consistency loss in semi-supervised deep learning classification methods, we visualized the semantic structure correlation matrices for different epochs. As shown in Figure 5, in the early stage of training, the correlations between samples are not well presented as the model has not yet converged, and the semantic structure correlation matrices of the student and teacher models under different perturbations show large differences. As training proceeded, the semantic correlations

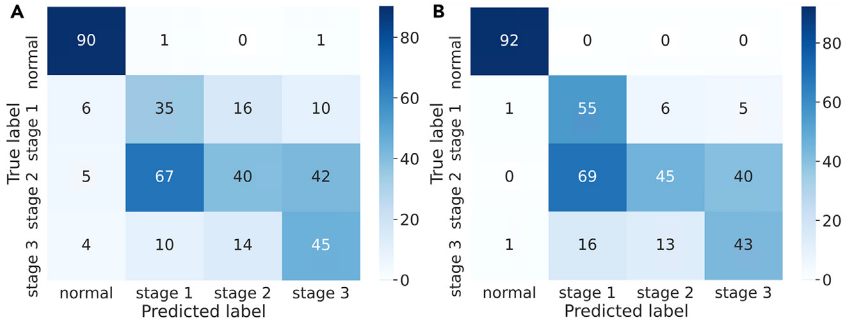


Figure 4. Confusion matrix of different methods

(A) Baseline 0.3.

(B) Ours.

Table 3. Classification performance of the model on the test set using different proportions of labeled data

	Percentage		Metrics			
Method	Labeled	Unlabeled	AUC	Accuracy	Sensitivity	Specificity
Oracle	100%	0%	0.8869	0.8225	0.7801	0.8359
Baseline_0.05	5%	95%	0.8027	0.6716	0.8301	0.6165
Ours	5%	95%	**0.8188**	**0.7442**	**0.8341**	**0.6919**
Baseline_0.1	10%	90%	0.7908	0.6988	0.7529	0.6787
Ours	10%	90%	**0.8059**	**0.7247**	**0.7550**	**0.7086**
Baseline_0.15	15%	85%	0.7960	0.6696	**0.7902**	0.5981
Ours	15%	85%	**0.8345**	**0.7655**	0.7687	**0.7622**
Baseline_0.2	20%	80%	0.8173	0.7169	**0.8470**	0.6497
Ours	20%	80%	**0.8283**	**0.7739**	0.7686	**0.7740**
Baseline_0.3	30%	70%	0.8226	0.7591	0.7562	0.7602
Ours	30%	70%	**0.8630**	**0.7895**	**0.7868**	**0.8010**
Baseline_0.5	50%	50%	0.8487	0.7979	0.7581	0.8016
Ours	50%	50%	**0.8751**	**0.8128**	**0.7754**	**0.8231**

Bold in the tables represent better performance results.

between the samples became clearer and the differences between the semantic structure correlation matrices of the student and teacher models became smaller due to the constraint of semantic structure correlation consistency loss, thus allowing the model to learn a more robust and discriminative feature representation, which explains the improved performance of the proposed method.

Heatmap visualization

To more visually demonstrate which regions of the image our method uses for prediction, we visualize the heatmap using Class Activation Mapping (CAM) techniques. As shown in Figure 6, it can be seen that our method mainly uses the features of the ridge region in the ROP image to determine which stage of ROP the image belongs to, which is consistent with the clinical experience of doctors in determining which stage the ROP image belongs to.

DISCUSSION

ROP is a retinal disease that affects premature newborns and causes blindness worldwide. Automated screening and diagnosis of ROP help doctors to give appropriate treatment plans in a timely and appropriate manner. A number of deep learning algorithms have been proposed, and significant progress has been made in order to implement automated ROP detection in real-life scenarios. For example, Tong et al.[27] used deep learning to perform automatic assessment of ROP severity and also to detect the presence of plus disease; the accuracy of its classification of ROP severity was 0.903. Mulay et al.[28] used a Mask region-based convolutional neural network (R-CNN)[29] model based on deep convolution neural networks to detect the important disease landmark ridge in ROP images, thus helping to better diagnose and screen for ROP early, and their model reached detection accuracy of 0.88. However, for real-world automatic ROP staging, obtaining large amounts of labeled data is often very difficult, which limits the application of deep learning models. This is mainly due to the following reasons. Firstly, the disease features of ROP are not very obvious at the beginning and the ROP images are usually acquired with low contrast, which leads to a lot of time and effort required by the doctor for annotation. Secondly, it is difficult to collect a large amount of ROP image data because the number of ROP patients is relatively small compared with other common disease. In this paper, we propose a new semi-supervised deep learning framework that aims to improve the classification performance of the model for ROP staging using a small amount of labeled data and a large amount of unlabeled data. We base our approach on a consistency regularization strategy, which encourages the model to be consistent in its predictions for data after different perturbations to make efficient use of the large amount of unlabeled data. In addition, we also consider that the disease evolution relationship between different ROP stages may be useful for classification performance and therefore further propose a semantic structure correlation consistency loss to encourage consistency in semantic structure across samples. Our approach was able to significantly reduce the annotation burden on doctors. As can be seen in Table 3, our approach achieves near fully supervised performance using only 50% of the data, which validates the effectiveness of the proposed approach. In addition, it is evident from the confusion matrix and receiver operating characteristic curves that our method performs better on all categories. The heatmap also demonstrates that our method does make judgments based on the shape and size of the important ROP-related disease feature ridge. Our method can be used not only for ROP staging but also for other computer-aided medical image analysis tasks and thus has significant clinical applications.

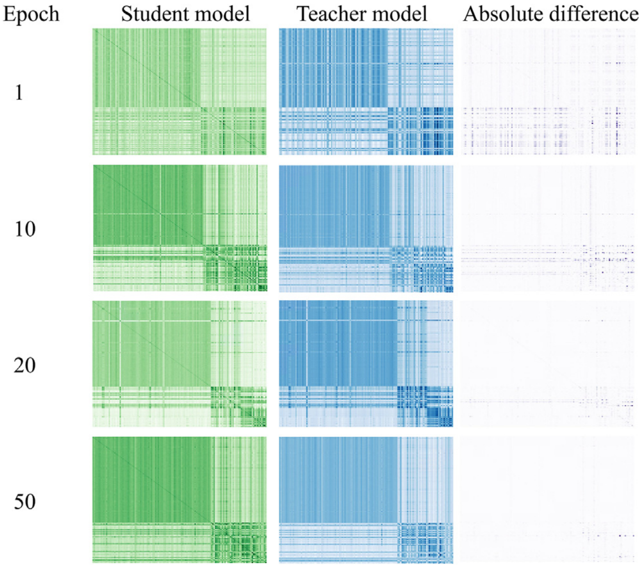


Figure 5. Evolution of the semantic structure correlation matrix.

There are several limitations of our work. Limited by the data source, our model has not been tested for performance on the ROP fundus image dataset from a large range of healthcare institutions. In addition, quality control of our dataset is done by experienced professionals. However, there may be some ROP images of poor quality in large-scale screening scenarios, which may affect the performance of the model.

Conclusion

In this paper, we propose a semi-supervised deep learning classification model for ROP staging. We propose two types of consistency loss to efficiently exploit unlabeled data, which can substantially reduce the annotation burden on doctors and require them to annotate only part of the data to achieve near fully supervised performance. We experimentally validate the effectiveness of the proposed algorithm. Future work is needed to validate the performance of the proposed algorithm on a larger medical clinical dataset.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability

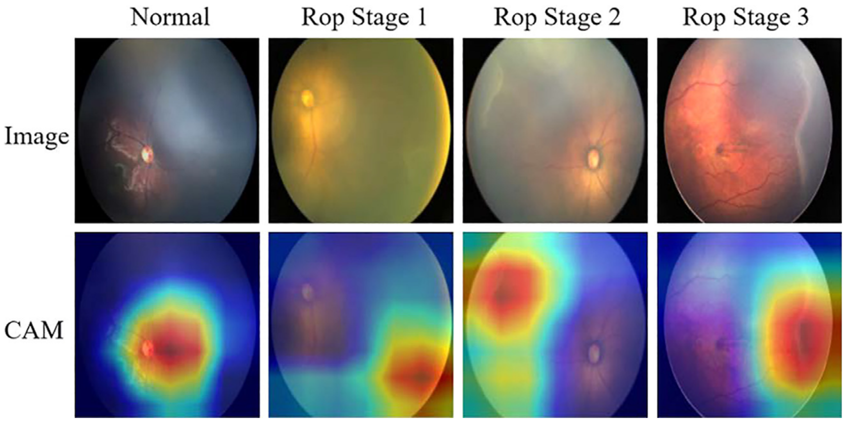


Figure 6. CAM heatmap visualization.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (81900908, 81770964); Science and Technology Commission of Shanghai Municipality, Shanghai Sailing Program (19YF1432500); and Shanghai Hospital Development Center (SHDC2020CR5014-002).

AUTHOR CONTRIBUTIONS

Q.H. and P.Z. obtained the dataset for the study and conducted the initial experiments. W.F., Q.H., T.M., L.J., Z.G., and Y.C. were involved in the revision of the study objectives and methods. All authors were involved in editing and proofreading. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

Authors W.F., L.J., T.M., and Y.C. were employed by Beijing Airdoc Technology Co., Ltd., China..

Received: March 20, 2023
Revised: July 3, 2023
Accepted: November 20, 2023
Published: November 23, 2023

REFERENCES

1. Gupta, K., Campbell, J.P., Taylor, S., Brown, J.M., Ostmo, S., Chan, R.V.P., Dy, J., Erdogmus, D., Ioannidis, S., Kalpathy-Cramer, J., et al. (2019). A quantitative severity scale for retinopathy of prematurity using deep learning to monitor disease regression after treatment. JAMA Ophthalmol. *137*, 1029–1036.
2. Taylor, S., Brown, J.M., Gupta, K., Campbell, J.P., Ostmo, S., Chan, R.V.P., Dy, J., Erdogmus, D., Ioannidis, S., Kim, S.J., et al. (2019). Monitoring disease progression with a quantitative severity scale for retinopathy of prematurity using deep learning. JAMA Ophthalmol. *137*, 1022–1028.
3. Redd, T.K., Campbell, J.P., Brown, J.M., Kim, S.J., Ostmo, S., Chan, R.V.P., Dy, J., Erdogmus, D., Ioannidis, S., Kalpathy-Cramer, J., and Chiang, M.F. (2019). Evaluation of a deep learning image assessment system for detecting severe retinopathy of prematurity. Br. J. Ophthalmol. *103*, 580–584.
4. Ting, D.S., Wu, W.-C., and Toth, C. (2019). Deep Learning for Retinopathy of Prematurity Screening.
5. Chan, T.H., Jia, K., Gao, S., Lu, J., Zeng, Z., and Ma, Y. (2015). PCANet: A Simple Deep Learning Baseline for Image Classification? IEEE Trans. Image Process. *24*, 5017–5032.
6. Yang, X., Ye, Y., Li, X., Lau, R.Y.K., Zhang, X., and Huang, X. (2018). Hyperspectral image classification with deep learning models. IEEE Trans. Geosci. Remote Sens. *56*, 5408–5423.
7. Young, T., Hazarika, D., Poria, S., and Cambria, E. (2018). Recent trends in deep learning based natural language processing. IEEE Comput. Intell. Mag. *13*, 55–75.
8. Wu, S., Roberts, K., Datta, S., Du, J., Ji, Z., Si, Y., Soni, S., Wang, Q., Wei, Q., Xiang, Y., et al. (2020). Deep learning in clinical natural language processing: a methodical review. J. Am. Med. Inform. Assoc. *27*, 457–470.
9. Zhang, Z., Geiger, J., Pohjalainen, J., Mousa, A.E.-D., Jin, W., and Schuller, B. (2018). Deep learning for environmentally robust speech recognition: An overview of recent developments. ACM Trans. Intell. Syst. Technol. *9*, 1–28.
10. Huang, J., and Kingsbury, B. (2013). Audio-visual deep learning for noise robust speech recognition. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE), pp. 7596–7599.
11. Peng, Y., Zhu, W., Chen, Z., Wang, M., Geng, L., Yu, K., Zhou, Y., Wang, T., Xiang, D., Chen, F., and Chen, X. (2021). Automatic staging for retinopathy of prematurity with deep feature fusion and ordinal classification strategy. IEEE Trans. Med. Imaging *40*, 1750–1762.
12. Wang, J., Ju, R., Chen, Y., Zhang, L., Hu, J., Wu, Y., Dong, W., Zhong, J., and Yi, Z. (2018). Automated retinopathy of prematurity screening using deep neural networks. EBioMedicine *35*, 361–368.
13. Lei, B., Zeng, X., Huang, S., Zhang, R., Chen, G., Zhao, J., Wang, T., Wang, J., and Zhang, G. (2021). Automated detection of retinopathy of prematurity by deep attention network. Multimed. Tools Appl. *80*, 36341–36360.
14. Laine, S., and Aila, T. (2016). Temporal ensembling for semi-supervised learning. Preprint at arXiv. https://doi.org/10.48550/arXiv.1610.02242.
15. Tarvainen, A., and Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Adv. Neural Inf. Process. Syst. *30*.
16. Cheplygina, V., de Bruijne, M., and Pluim, J.P.W. (2019). Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med. Image Anal. *54*, 280–296.
17. Wang, D., Zhang, Y., Zhang, K., and Wang, L. (2020). Focalmix: Semi-supervised learning for 3d medical image detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3951–3960.
18. Chen, S., Bortsova, G., Garc'ıa-Uceda Jua'rez, A., Tulder, G.v., and Bruijne, M.d. (2019). Multi-task attention-based semi-supervised learning for medical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (Springer), pp. 457–465.
19. Shang, H., Sun, Z., Yang, W., Fu, X., Zheng, H., Chang, J., and Huang, J. (2019). Leveraging other datasets for medical imaging classification: evaluation of transfer, multi-task and semi-supervised learning. In International conference on medical image computing and computer-assisted intervention (Springer), pp. 431–439.
20. Bortsova, G., Dubost, F., Hogeweg, L., Katramados, I., and Bruijne, M.d. (2019). Semi-supervised medical image segmentation via learning consistency under transformations. In International Conference on Medical Image Computing and Computer-Assisted Intervention (Springer), pp. 810–818.
21. Zhang, W., Zhu, L., Hallinan, J., Zhang, S., Makmur, A., Cai, Q., and Ooi, B.C. (2022). Boostmis: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20666–20676.
22. Zhao, R., Chen, X., Liu, X., Chen, Z., Guo, F., and Li, S. (2020). Direct cup-to-disc ratio estimation for glaucoma screening via semi-supervised learning. IEEE J. Biomed. Health Inform. *24*, 1104–1113.
23. Adal, K.M., Sidibé, D., Ali, S., Chaum, E., Karnowski, T.P., and Mériaudeau, F. (2014).

Automated detection of microaneurysms using scale- adapted blob analysis and semi-supervised learning. Comput. Methods Programs Biomed. *114*, 1–10.

24. Liu, S., Hong, J., Lu, X., Jia, X., Lin, Z., Zhou, Y., Liu, Y., and Zhang, H. (2019). Joint optic disc and cup segmentation using semi-supervised conditional gans. Comput. Biol. Med. *115*, 103485.
25. Liu, Q., Yu, L., Luo, L., Dou, Q., and Heng, P.A. (2020). Semi-supervised medical image classification with relation-driven self-ensembling model. IEEE Trans. Med. Imaging *39*, 3429–3440.
26. Gatys, L.A., Ecker, A.S., and Bethge, M. (2015). A neural algorithm of artistic style. Preprint at arXiv. https://doi.org/10.48550/arXiv.1508.06576.
27. Tong, Y., Lu, W., Deng, Q.-q., Chen, C., and Shen, Y. (2020). Automated identification of retinopathy of prematurity by image-based deep learning. Eye Vis. *7*, 40.
28. Mulay, S., Ram, K., Sivaprakasam, M., and Vinekar, A. (2019). Early detection of retinopathy of prematurity stage using deep learning approach. In Medical Imaging 2019: Computer-Aided Diagnosis, *10950*Medical Imaging 2019: Computer-Aided Diagnosis (SPIE), pp. 758–764.
29. He, K., Gkioxari, G., Doll’ar, P., and Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision, pp. 2961–2969.
30. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
Resnet	He et al.[30] (2016)	https://doi.org/10.48550/arXiv.1512.03385
Mean teacher	Tarvainen et al.[15] (2017)	https://doi.org/10.48550/arXiv.1703.01780
Matplotlib	Version 3.3.1	https://matplotlib.org/3.3.1/
Scikit-learn	Version 0.23.2	https://scikit-learn.org/stable/whats_new/v0.23
Python	Version 3.80	https://www.python.org/downloads/release/python-380/
PyTorch	Version 1.11.0	https://download.pytorch.org/whl/torch/
Semi-supervised deep learning classification model	This study	Please request from lead contact (hqj1010@126.com) for non-commercial, research purposes

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Qiujing Huang (e-mail: hqj1010@126.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

(1) All data reported in this paper will be shared by the lead contact upon request.
(2) This paper does not report original code.
(3) Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Ethical statement

The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The image data were collected and analyzed with the approval of NCH and in compliance with the tenets of the Declaration of Helsinki.

METHOD DETAILS

Patient cohorts

The training dataset incorporated fundus images from 473 infants at Xinhua Hospital, maintaining a gender ratio of 1.2:1 (male: female). On average, the gestational age was (29.86 ± 2.47) weeks, and the mean birth weight was (1379.95 ± 412.95) grams. Within this dataset, there were 4796 fundus images, consisting of 3171 normal cases, 744 instances of stage 1 ROP, 426 instances of stage 2 ROP, and 455 instances of stage 3 ROP. For evaluating the model's performance, a separate test set was created using fundus images from 62 infants, also sourced from Xinhua Hospital. The gender ratio in the test set was 1.4:1 (male: female), with an average gestational age of (29.29 ± 2.94) weeks and a mean birth weight of (1381.77 ± 451.75) grams. This test set comprised 386 fundus images, including 92 normal cases, 67 stage 1 ROP cases, 154 stage 2 ROP cases, and 73 stage 3 ROP cases.

Semi-supervised deep learning classification model

In the realm of medical imaging, annotating medical images is an arduous and time-consuming task. The intricacies of medical conditions, structures, and anomalies require meticulous labeling, often demanding significant expertise from skilled professionals. The exhaustive nature of this annotation process poses a considerable challenge, as it necessitates a substantial investment of time and human resources.

Moreover, the development and training of deep learning models for medical image analysis heavily rely on vast amounts of meticulously annotated data. The hunger for labeled datasets is insatiable in the pursuit of creating robust and accurate models. The scarcity of annotated medical images, coupled with the intricate nature of medical phenomena, exacerbates the already formidable challenge of obtaining the requisite labeled data for effective model training.

To alleviate the annotation requirements and improve the ROP classification performance, we propose a semi-supervised learning-based classification framework. Our proposed semi-supervised deep learning classification model shares similarities with the well-known semi-supervised classification model, Mean Teacher. It comprises a student model and a teacher model, both identical in structure. The teacher model's parameters are optimized using the exponential moving average algorithm applied to the parameters of the student model. To optimize the student model's parameters, we employ supervised cross-entropy loss on labeled data and consistency loss on unlabeled data. The consistency loss is computed based on the outputs of both the student and teacher models and includes prediction consistency loss and semantic structure consistency loss. Prediction consistency loss enhances model generalization by ensuring both models yield the same predictions for the same input image with different data augmentations, extracting information from unlabeled images. We employ various techniques for introducing random perturbations in the data, including random horizontal/vertical flipping, random rotation, and color jittering. Additionally, a dropout layer with a dropout rate of 0.2 is incorporated before the global pooling layer of the ResNet50 network. Additionally, considering the semantic relevance between different stages of the ROP, we leverage semantic structure consistency loss to encourage the model to maintain consistent semantic structures for fundus images after data augmentation. This further extracts valuable information from unlabeled images, contributing to improved classification performance. We compute the case-wise gram matrix between samples, which reflects the semantic structural relationships between different samples. To standardize the input, all fundus images are resized to 224 × 224 for input to the classification model. Resnet50 serves as our network backbone, initialized with pretrained weights from the ImageNet dataset. Both the student and teacher models share an identical ResNet50 structure. The experiments were conducted using Python (3.8), PyTorch (1.11.0) and scikit-learn (0.23.2), and all experiments were executed on two Nvidia 3090 Ti GPUs.

QUANTIFICATION AND STATISTICAL ANALYSIS

We evaluate the performance of the proposed semi-supervised classification framework on the test set data from Xinhua Hospital. We employ AUC (Area Under the Curve), Accuracy, Sensitivity, and Specificity as the evaluation metrics to assess the performance of diverse algorithms. For heatmap visualization, we use the CAM technique, which helps us to understand on which features in the image the model is based to make predictions.