Deep Semi-Supervised and Self-Supervised Learning for Diabetic Retinopathy Detection

Jose Arrieta\textsuperscript{a}, Oscar J. Perdomo\textsuperscript{b}, and Fabio A. González\textsuperscript{a}

\textsuperscript{a}MindLab Research Group, Universidad Nacional de Colombia, Bogotá, Colombia; 
\textsuperscript{b}Universidad del Rosario, Bogotá, Colombia;

ABSTRACT

Diabetic retinopathy (DR) is one of the leading causes of blindness in the working-age population of developed countries, caused by a side effect of diabetes that reduces the blood supply to the retina. Deep neural networks have been widely used in automated systems for DR classification on eye fundus images. However, these models need a large number of annotated images. In the medical domain, annotations from experts are costly, tedious, and time-consuming; as a result, a limited number of annotated images are available. This paper presents a semi-supervised method that leverages unlabeled images and labeled ones to train a model that detects diabetic retinopathy. The proposed method uses unsupervised pretraining via self-supervised learning followed by supervised fine-tuning with a small set of labeled images and knowledge distillation to increase the performance in a classification task. This method was evaluated on the EyePACS test and Messidor-2 dataset achieving 0.94 and 0.89 AUC respectively using only 2% of EyePACS train labeled images.

Keywords: Diabetic retinopathy, medical imaging, deep learning, semi-supervised learning, self-supervised learning.

1. INTRODUCTION

Diabetic retinopathy (DR) is one of the leading causes of blindness in the working-age population in developed countries.\textsuperscript{1} DR consists of a side effect of diabetes that reduces the blood supply to the retina, including lesions that appear on the surface, such as microaneurysms, exudates, hemorrhages, and cotton wool spots. Automated methods for detecting eye diseases present as a useful tool for early diagnosis, important to prevent the occurrence of blindness and lack of vision.\textsuperscript{2,3} Recently, deep learning methods became popular for DR classification on eye fundus images because of their promising results.\textsuperscript{4–11} However, the majority of these approaches use labeled images, whereas the manual labeling of medical images is expensive and time-consuming because it requires medical experts in the retina. As a result, relatively small labeled data sets are available to train deep learning models.

Some authors have proposed semi-supervised learning approaches to leverage unlabeled images and mitigate the lack of annotated images. In specific, Liu et al.\textsuperscript{12} and Xie et al.\textsuperscript{13} propose the use of generative adversarial networks (GAN) to extend classification while performing unsupervised image reconstruction with unlabeled data in training. State-of-the-art methods in deep semi-supervised methods such as MixMatch\textsuperscript{14} and FixMatch\textsuperscript{15} comprises strong and weak augmentations of unlabeled images and makes class pseudo-labels match via consistency regularization. In medical imaging, Hansen et al.\textsuperscript{16} explore MixMatch in two distinct medical imaging domains as skin lesion diagnosis and lung cancer prediction. Pooch et al.\textsuperscript{17} evaluate the performance of different methods including pseudo-labeling, Mean Teacher,\textsuperscript{18} Unsupervised Data Augmentation (UDA),\textsuperscript{19} MixMatch\textsuperscript{14} and FixMatch\textsuperscript{15} in a chest radiography classification task.

Alternatively, self-supervised learning\textsuperscript{20} presents a new strategy to use unlabeled data to pre-train a neural network and allows the construction of models that can learn relevant image representations from unlabeled

Further author information: (Send correspondence to M.P.)
J.A.: E-mail: jmarrietar@unal.edu.co
O.J.P.: E-mail: oscar.j.perdomo@urosario.edu.co
F.A.G.: E-mail: fagonzalezo@unal.edu.co
The proposed method comprises four stages as follows: (1) Self-supervised pre-training using SimCLR with unlabeled images. (2) Supervised fine-tuning on a small set of labeled images. (3) Knowledge distillation from teacher model to student model using teacher’s output as pseudo-labels to train student model. (4) Fine-tuning of student model on a small set of labeled images.

medical images. These visual representations from unlabeled data are used to train strong semi-supervised models. Azizi et al. presented the effectiveness of self-supervised learning with unlabeled data for domain-specific medical images in dermatology skin and chest X-ray classification. Vu et al. use self-supervised contrastive learning and propose to use patient metadata to select the pairs of medical images requiring them to be from the same patient. Kaku et al. minimize the mean squared error between the intermediate layer representations in complement to the contrastive loss using medical datasets such as NIH-Chest X-rays, breast cancer histopathology, and diabetic retinopathy.

This paper presents a method that combines self-supervised learning along with a teacher-student and knowledge distillation strategies to perform semi-supervised learning with a small subset of labeled eye fundus images. The method achieves an AUC of 0.94 on the EyePACS test and an AUC of 0.89 on the Messidor-2 Diabetic Retinopathy datasets using only 2% EyePACS-Kaggle train labeled images (1000 images).

2. METHODS

The proposed method consists of a teacher and a student network. Our method differs from previous approaches because the teacher network heavily relies on large unlabeled diabetic retinopathy (DR) images to learn useful representations in pre-training compared to the labeled ones available. Furthermore, the student network benefits from pseudo-labels created with the teacher model using unlabeled images. The complete semi-supervised workflow proposed is summarized in Figure 1 and consists of four stages.

1) Self-supervised unlabeled pre-training: First; the Teacher model is pre-trained using the simple framework for contrastive learning of visual representation (SimCLR). This process is unsupervised as it doesn’t need the labels of the images. The whole EyePACS-Kaggle train dataset (57146 images) without the labels and a ResNet-50 backbone were used to learn useful visual representations with domain-specific medical images. The SimCLR approach uses strong data augmentations such as color distortions (brightness, contrast, saturation, hue), cropping, and rotations as shown in Figure 2. Contrastive loss $NT-Xent$ (the normalized temperature-scaled cross-entropy loss) is used to maximize the agreement of diverse representations of the same image in a
latent space, defined as:

$$L_{i,j} = -\log \frac{\exp \left( \frac{\text{sim} \left( z_i, z_j \right)}{\tau} \right)}{\sum_{k=1}^{2N} \exp \left( \frac{\text{sim} \left( z_i, z_k \right)}{\tau} \right)}$$ (1)

where $z_i$ and $z_j$ correspond to the outputs of the same image from the non-linear projection head and $z_k$ corresponds to the remainder of representations of the other $N$ images in the batch. Additionally, $\text{sim}$ denotes a distance metric and $\tau$ the temperature parameter to control the softness of distribution.

2) **Fine-tuning of Teacher model with labeled images**: In this phase, a small set of 1000 random samples of labeled fundus images are used to further refine the weights of the model. Fine-tuning is a common approach where a network from a previous task is used as starting point, and weights are adjusted to fit a new specific task. The proposed approach used the previously pre-trained SimCLR ResNet-50$^{25}$ model with an additional sigmoid layer at the top to make a binary classification from referable DR and non-referable DR.

3) **Pseudo-labeling and knowledge distillation**: The third step consists of knowledge distillation from the ResNet-50 model (Teacher) to a DenseNet161 model (Student) using unlabeled data as input and teacher’s output as pseudo-labels. Knowledge distillation is the process of transferring knowledge from one model to another one, originally used to compress a large model to a smaller one, encouraging the student to match a
However, we chose a DenseNet161, a deeper network than the teacher for Student architecture because Xie et al. demonstrated that using student models that are equal to or larger than the teacher in addition to some noise to the student can improve the performance of the student beyond the teacher, making predictions with more difficult images by giving the student model enough capacity and difficult images to learn through. We transform soft-pseudo-labels into hard-pseudo-labels using a threshold of 0.5.

4) Fine-tuning of Student model with labeled images: Finally, the fourth task consists of a final fine-tuning of the student model with the same set of 1000 labeled eye fundus images.

3. EXPERIMENTAL EVALUATION

3.1 Datasets

EyePACS-Kaggle consists of retina images taken under a variety of imaging conditions as shown in figure 3, where the clinician rate the presence of diabetic retinopathy in each image on a scale of 0 to 4 (No DR, mild, moderate, severe, and proliferative), respectively. For the binary classification task, grades 0 and 1 correspond to non-referable DR, while grades 2, 3, and 4 correspond to referable DR according to the International Clinical Diabetic Retinopathy Scale. The EyePACS-Kaggle dataset is partitioned into train and test datasets, with 35121 images and 42918 images respectively of different sizes. In this work, we used the EyePACS-Kaggle partition defined in Voets et al. which consists of a train set of 57146 images and a test set of 8790 images. Training is performed on the EyePACS-Kaggle train and evaluation is performed on the EyePACS-Kaggle test and Messidor-2 datasets. Messidor-2 consists of a public dataset that contains 1748 images with grades adjudicated by a panel of three retina specialists and constitutes a standard dataset used to compare performance results in Diabetic Retinopathy detection. Following Voets et al. and Gulshan et al. pre-processing, all fundus images are centered and resized to 299 x 299 pixels, with the fundus center in the middle of the image.

![Figure 3. Example of and Eye Fundus Image (EFI) non-referable DR on the left and a referable DR example on the right.](image)

3.2 Experimental setup

All models were implemented using Pytorch. EyePACS-Kaggle train dataset without labels is used for self-supervised pretraining a ResNet-50 network using SimCLR data augmentations alongside a projection head with a contrastive loss function, a learning rate of $1.0 \times 10^{-5}$, weight decay of $5.0 \times 10^{-4}$, and a batch of size 64 for 100 epochs. For knowledge distillation, the previously trained ResNet-50 network is used as a teacher to predict pseudo-labels of the EyePACS-Kaggle train dataset while using them to train a different architecture. DenseNet161, a deeper network is used to improve the performance of the student beyond the teacher. The data augmentation configuration consists of random color jitter and random horizontal flips with BinaryCrossEntropy loss function and stochastic gradient descent as an optimizer, a learning rate of $1.0 \times 10^{-4}$, and a batch size 32 for 200 epochs. A final fine-tuning of the student network is performed using the 1000 images labeled data for 100 epochs.

Three baselines were defined. First, the supervised Inception-V3 network as proposed in Voets et al. and Krause et al., although some other architectures were also tested with similar performance. Additionally, two state-of-the-art hybrid methods, such as MixMatch and FixMatch, were chosen to compare the performance of
Table 1. AUC performance measures of baselines and proposed method on test datasets. A supervised baseline is trained using 2% of EyePACS train labeled images. Semi-supervised methods use 2% of EyePACS train labeled images and the rest as unlabeled images.

| Label fraction | Method                          | Architecture | Kaggle Test | Messidor2 |
|----------------|---------------------------------|--------------|-------------|-----------|
| 100%           | Supervised                      | InceptionV3  | 0.96        | 0.88      |
| 2% Train       | Supervised                      | InceptionV3  | 0.88        | 0.66      |
|                | MixMatch\textsuperscript{14}    | ResNet50     | 0.85        | 0.64      |
|                | FixMatch\textsuperscript{15}    | ResNet50     | 0.83        | 0.79      |
|                | SimCLR-Finetuned (Teacher)      | ResNet50     | 0.92        | 0.85      |
|                | SimCLR-Distilled (Student)      | DenseNet161  | 0.94        | 0.89      |

Figure 4. t-SNE visual representation in Messidor-2: (a) Supervised, (b) SimCLR-Finetuned (Teacher), and (c) SimCLR-Distilled methods (Student).

semi-supervised approaches. The metric used for comparison is AUC, as this is a binary classification problem with imbalance classes and is a common metric used in related papers.

4. RESULTS

The proposed method is compared to a supervised baseline that uses the fraction of 1000 labeled images (2% of EyePACS-Kaggle) and semi-supervised learning methods such as FixMatch and MixMatch that make class pseudo-labels match via consistency regularization using additional unlabeled data. Additionally, a supervised method with the complete train dataset is used as a reference for performance. The proposed method successfully outperforms the methods trained with a labeled fraction of 2% of EyePACS train and its at a similar performance compared to 100% of the train labeled images method leveraging the use of unlabeled data in the pre-train stage and creating useful visual representations. Additionally, the knowledge distillation step creates a gain under the AUC metric as reported in Table 1 completes a better generalization task of detection of diabetic retinopathy using only 1000 labeled images 2% of EyePACS-Kaggle and the remainder as unlabeled images. Further, in Figure 4 the representation of visual features is presented using the t-SNE technique in the Messidor-2 dataset. The supervised representation of DR and non-DR seems not easily separable. In contrast, the SimCLR-Finetuned (Teacher) and SimCLR-Distilled (Student) seems more separable between the DR and non-DR instances, improving visual representations and generalization.

5. DISCUSSION AND CONCLUSION

Despite classical deep learning models requiring a large amount of labeled data for their training process, this paper presents a method that combines self-supervised learning along with teacher-student and knowledge distillation strategies to perform semi-supervised learning with a small subset of labeled eye fundus images. Our method achieved an AUC of 0.94 tested on the EyePACS test dataset, and an AUC of 0.89 tested on the Messidor-2 Diabetic Retinopathy dataset using only 2% of EyePACS train labeled images that improved the
performance from the baseline supervised approach, and with greater performance than the other two state-of-the-art methods, alongside a better generalization compared to the teacher model. Self-supervised learning resulted in a great pre-training strategy to use unlabeled medical images and create richer and better visual representations improving downstream tasks such as diabetic retinopathy detection in this case. Although one disadvantage of this approach consists of computing and memory intensive, for this case 48 hours approx as it creates several representations from each image in the training batch and could require a lengthy training process if more images are added. Also, knowledge distillation from a teacher to a student model improved performance in the presented results. Furthermore, the student network benefits from pseudo-labels created with the teacher model using unlabeled images, improving the generalization of outcomes and enhancing the performance beyond the teacher. In future works, we will propose the validation of our model in a clinical environment, including images from different sources and multiple ocular datasets such as APTOS 2019, DeepDRiD, and IDRiD. Additionally, an ablation study for different label fractions (1%, 5%, 10%) of the training set to analyze gains for the student and teacher models. Finally, we hypothesized that this method could be applied to other medical image domains where labeled images are scarce and constitute a vital research topic for deep learning models.

REFERENCES

[1] Organisation, W. H., [World report on vision. 2019], vol. 214 (2019).
[2] Bhaskaranand, M., Ramachandra, C., Bhat, S., Cuadros, J., Nittala, M. G., Sadda, S. R., and Solanki, K., “The value of automated diabetic retinopathy screening with the EyeArt system: A study of more than 100,000 consecutive encounters from people with diabetes,” Diabetes Technology and Therapeutics 21(11), 635–643 (2019).
[3] Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Raunviboonsuk, D. P., and Vardoulakis, L., “A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy,” CHI 2020 Paper, 1–12 (2020).
[4] Asiri, N., Hussain, M., Al, F., and Alzaidi, N., “Deep learning based computer-aided diagnosis systems for diabetic retinopathy: A survey,” Artificial Intelligence In Medicine 99 (December 2018), 101701 (2019).
[5] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., and Webster, D. R., “Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs,” 94043, 1–9 (2016).
[6] Voets, M., Kajsa, M., and Bongo, L., “Reproduction study using public data of: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs,” 1–11 (2019).
[7] Krause, J., Gulshan, V., Rahimy, E., Karth, P., Widner, K., Corrado, G. S., Peng, L., and Webster, D. R., “Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy,” Ophthalmology 125(8), 1264–1272 (2018).
[8] Quellec, G., Charrière, K., Boudi, Y., and Cochener, B., “Deep image mining for diabetic retinopathy screening,” 39, 178–193 (2017).
[9] Rakhlin, A., “Diabetic Retinopathy detection through integration of Deep Learning classification framework,” 1–11 (2018).
[10] Toledo-Cortés, S., De La Pava, M., Perdómo, O., and González, F. A., “Hybrid deep learning gaussian process for diabetic retinopathy diagnosis and uncertainty quantification,” 206–215 (2020).
[11] delaPava, M., Rios, H., Rodríguez, F., Perdomo, O., and González, F., “A deep learning model for classification of diabetic retinopathy in eye fundus images based on retinal lesion detection,” in [Proc. of SPIE], 12088, 120880T.
[12] Liu, S., B, J. X., Wu, J., and Shi, P., “Semi-supervised Adversarial Learning for Diabetic Retinopathy Screening,” 60–68 (2019).
[13] Xie, Y., Wan, Q., Chen, G., Xu, Y., and B, B. L., [Retinopathy Diagnosis Using Generative Adversarial Network], vol. 1, Springer International Publishing.
[14] Berthelot, D., Oliver, A., Carlini, N., Goodfellow, I., Raffel, C., and Papernot, N., “MixMatch : A Holistic Approach to Semi-Supervised Learning,” (NeurIPS), 1–11 (2019).
Sohn, K., Berthelot, D., Zizhao, C.-l. L., Nicholas, Z., Cubuk, E. D., Kurakin, A., Zhang, H., and Raffel, C., “FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12446 LNCS, 112–121 (2020).

Hansen, C. B., Nath, V., Gao, R., Bermudez, C., Huo, Y., Sandler, K. L., Massion, P. P., Blume, J. D., Lasko, T. A., and Landman, B. A., “Semi-supervised Machine Learning with MixMatch and Equivalence Classes,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12446 LNCS, 172–179 (2020).

Pooch, E. H., Ballester, P., and Barros, R. C., “Semi-supervised Classification of Chest Radiographs,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12446 LNCS, 112–121 (2020).

Tarvainen, A. and Valpola, H., “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” Advances in Neural Information Processing Systems 2017-December, 1196–1205 (2017).

Xie, Q., Dai, Z., Hovy, E., Luong, M.-t., and Le, Q. V., “Unsupervised Data Augmentation for Consistency Training,” (NeurIPS), 1–20 (2020).

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G., “A Simple Framework for Contrastive Learning of Visual Representations,” (2019).

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G., “Big Self-Supervised Models are Strong Semi-Supervised Learners,” (NeurIPS), 1–18 (2020).

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., Nataraajan, V., and Norouzi, M., “Big Self-Supervised Models Advance Medical Image Classification,” 1–19.

Vu, Y. N. T., Wang, R., Balachandar, N., Liu, C., Ng, A. Y., and Rajpurkar, P., “MedAug: Contrastive learning leveraging patient metadata improves representations for chest X-ray interpretation,” 1–14 (2021).

Kaku, A., Upadhyay, S., and Razavian, N., “Intermediate layers matter in momentum contrastive self-supervised learning,” (2021).

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K., “Aggregated residual transformations for deep neural networks,” Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 2017-Janua, 5987–5995 (2017).

Hinton, G., Vinyals, O., and Dean, J., “Distilling the Knowledge in a Neural Network,” 1–9 (2015).

retinopathy detection, D., “Eyepacs.” www.kaggle.com/competitions/diabetic-retinopathy-detection (2015).

Scale, I. C. D. R. D. S., “International Clinical Diabetic Retinopathy,” (October), 8500 (2002).

de Traitement de l’Information Médicale, L., “Messidor-2 dataset.” https://www.adcis.net/en/third-party/messidor2/ (2011).

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S., “Pytorch: An imperative style, high-performance deep learning library,” in Advances in Neural Information Processing Systems 32, 8024–8035, Curran Associates, Inc. (2019).