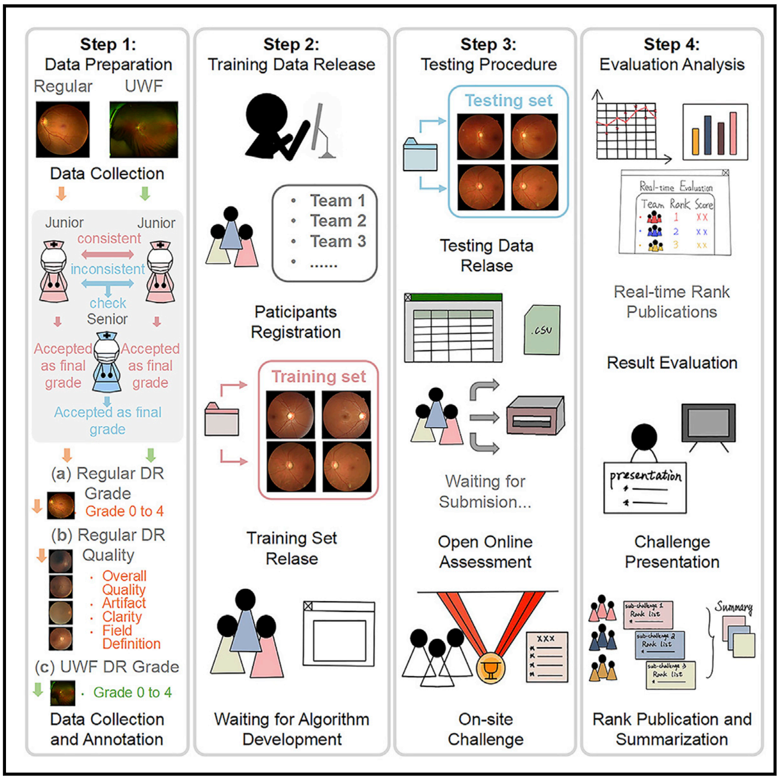
Descriptor

Patterns

DeepDRiD: Diabetic Retinopathy—Grading and Image Quality Estimation Challenge

Graphical abstract



Highlights

- Provides the DeepDRiD dataset, performance evaluation, top methods and results
- Presents deep learning approaches in DR image quality assessment and grading
- Discusses the future work of DR automatic screening

Authors

Ruhan Liu, Xiangning Wang, Qiang Wu, ..., Dinggang Shen, Bin Sheng, Ping Zhang

Correspondence

huarting99@sjtu.edu.cn (H.L.),
dinggang.shen@gmail.com (D.S.),
shengbin@sjtu.edu.cn (B.S.)

In brief

In DeepDRiD challenge, organizers hold a real-world exploration in diabetic retinopathy (DR) auto-screening systems using regular fundus images from 500 participants and ultra-widefield fundus images from 128 participants. Among the 34 participating teams, we summarized the top 3 teams in the three sub-challenges involved in DR grading and image quality assessment. In addition to providing new insights into image quality assessment strategy, these models can enhance the judgment of healthcare workers in DR screening and bring precise screening results.

Liu et al., 2022, Patterns *3*, 100512
June 10, 2022 © 2022 The Author(s).
https://doi.org/10.1016/j.patter.2022.100512

Descriptor

DeepDRiD: Diabetic Retinopathy—Grading and Image Quality Estimation Challenge

Ruhan Liu,[1,2,25] Xiangning Wang,[3,25] Qiang Wu,[3,25] Ling Dai,[1,2] Xi Fang,[4] Tao Yan,[5] Jaemin Son,[6] Shiqi Tang,[7] Jiang Li,[8] Zijian Gao,[9] Adrian Galdran,[10] J.M. Poorneshwaran,[11] Hao Liu,[9] Jie Wang,[12] Yerui Chen,[13] Prasanna Porwal,[14]

(Author list continued on next page)

[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[2]MoE Key Lab of Artificial Intelligence, Artificial Intelligence Institute, Shanghai Jiao Tong University, Shanghai, China
[3]Department of Ophthalmology, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai, China
[4]Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China
[5]Department of Electromechanical Engineering, University of Macau, Macao, China
[6]VUNO Inc., Korea
[7]Department of Mathematics, City University of Hong Kong, Hong Kong, China
[8]Institute of Image Processing and Pattern Recognition, Department of Automation, Shanghai Jiao Tong University, Shanghai, China
[9]School of Electronic Information, Hangzhou Dianzi University, Hangzhou, China

(Affiliations continued on next page)

THE BIGGER PICTURE Diabetic retinopathy (DR) is the most common disease caused by diabetes. Challenges are held to address real-world issues encountered in the design of DR automated screening systems to advance the technology in this area. Thus, we described a challenge named "Diabetic Retinopathy (DR)—Grading and Image Quality Estimation Challenge" in conjunction with the IEEE International Symposium on Biomedical Imaging (ISBI 2020) for fundus image assessment and DR grading. The scientific community responded positively to the challenge. In the challenge, we provided a deep DR image dataset (DeepDRiD) containing regular DR images and ultra-widefield (UWF) DR images, both having image quality and DR grading diagnosis. We discussed details of the three best algorithms in each sub-challenges. The results by the top algorithms showed that image quality assessment can be used as a target for further exploration.

1 2 3 4 5 Proof-of-concept Data science output has been formulated, implemented, and tested for one domain/problem

SUMMARY

We described a challenge named "Diabetic Retinopathy (DR)—Grading and Image Quality Estimation Challenge" in conjunction with ISBI 2020 to hold three sub-challenges and develop deep learning models for DR image assessment and grading. The scientific community responded positively to the challenge, with 34 submissions from 574 registrations. In the challenge, we provided the DeepDRiD dataset containing 2,000 regular DR images (500 patients) and 256 ultra-widefield images (128 patients), both having DR quality and grading annotations. We discussed details of the top 3 algorithms in each sub-challenges. The weighted kappa for DR grading ranged from 0.93 to 0.82, and the accuracy for image quality evaluation ranged from 0.70 to 0.65. The results showed that image quality assessment can be used as a further target for exploration. We also have released the DeepDRiD dataset on GitHub to help develop automatic systems and improve human judgment in DR screening and diagnosis.

INTRODUCTION

Diabetic retinopathy (DR) is the most common disease caused by diabetes, and it leads to vision loss in adults and mainly affects the working-age population.[1–4] Approximately 600 million people are estimated to have diabetes by 2040, and one-third of them are expected to have DR.[1] DR is diagnosed by visually inspecting a retinal fundus image for the presence of one or

Gavin Siew Wei Tan,[15] Xiaokang Yang,[2] Chao Dai,[16] Haitao Song,[2] Mingang Chen,[17] Huating Li,[18,19,*] Weiping Jia,[18,19] Dinggang Shen,[20,21,*] Bin Sheng,[1,2,25,26,*] and Ping Zhang,[22,23,24]

[10]Bournemouth University, United Kingdom
[11]Healthcare Technology Innovation Centre, IIT Madras, India
[12]School of Computer Science and Engineering, Beihang University, Beijing, China
[13]Nanjing University of Science and Technology, Nanjing, China
[14]Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded, India
[15]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore
[16]Shanghai Zhi Tang Health Technology Co., LTD., China
[17]Shanghai Key Laboratory of Computer Software Testing & Evaluating, Shanghai Development Center of Computer Software Technology, Shanghai, China
[18]Department of Endocrinology and Metabolism, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai, China
[19]Shanghai Diabetes Institute, Shanghai Clinical Center for Diabetes, Shanghai, China
[20]School of Biomedical Engineering, ShanghaiTech University, Shanghai, China
[21]Department of Research and Development, Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China
[22]Department of Computer Science and Engineering, The Ohio State University, Ohio, USA
[23]Department of Biomedical Informatics, The Ohio State University, Ohio, USA
[24]Translational Data Analytics Institute, The Ohio State University, Ohio, USA
[25]These authors contributed equally
[26]Lead contact
*Correspondence: huarting99@sjtu.edu.cn (H.L.), dinggang.shen@gmail.com (D.S.), shengbin@sjtu.edu.cn (B.S.)

https://doi.org/10.1016/j.patter.2022.100512

more retinal lesions, such as microaneurysms, hemorrhages, soft exudates, and hard exudates.[5] An internationally accepted method of grading the DR levels classifies DR into non-proliferative DR (NPDR) and proliferative DR (PDR).[4] NPDR is the early stage of DR and is characterized by the presence of microaneurysms, whereas PDR is an advanced stage of DR and can lead to severe vision loss. The number and degree of retinal lesions vary in different DR grading, and the specific grading standards of NPDR and PDR are listed in Table 1.[4] Furthermore, Figure S1 shows different levels of DR disease presentation.

DR has some severe implications, such as blindness, and its whole population screening is still hampered by several factors.[6–10] First, DR screening places a cumbersome burden on ophthalmologists. Second, healthcare workers are faced with inadequate training, resulting in low-accuracy problems in DR grading.[11] Therefore, computer-aided diagnostic tools are needed to assist manual screening, reducing the burden on ophthalmologists, and helping trained providers to grade fundus images more accurately.[12–16] Recent studies have been conducted to collect raw fundus images and achieve accurate pixel- or image-level expert annotations;[11,17–19] these efforts play an important role in facilitating the research community in developing, validating, and comparing DR gradings. Large numbers of raw fundus images and their corresponding physician annotations have important clinical implications for developing robust automated DR grading models.

In medical image analysis, grand competitions present substantial opportunities to quickly advance the state-of-the-art methods. Organizers define a clinically relevant task and build a sufficiently large and diverse dataset to allow participants to develop algorithms for solving one or several clinically related problem(s). Moreover, algorithms proposed by participants are consistently evaluated in a fair performance comparison. Many successful challenges have been organized in recent years, specifically in DR fields, i.e., IDRiD,[20] Kaggle 2015,[21] Messidor,[22] Kaggle 2009,[23] ROC,[24] E-Ophtha,[25] and DiaretDB.[26]

In 2018, we participated in the ''Diabetic Retinopathy—Segmentation and Grading Challenge'' (IDRiD) to grade DR levels, segment fundus lesions, and locate retinal landmarks (macula and optic disc) in regular fundus images.[20] The development of our automated screening system for DR was further refined during and after the competition. The details of the model development are described in Dai et al.'s work.[11] When developing the system, we found certain problems hindering the practicality of the automatic DR screening system. First, low-quality fundus images due to significant artifacts and poorly lit areas increase training difficulty. Moreover, fundus images from different devices pose a challenge for the stability of automated screening systems. Finally, dual views of regular fundus images are rarely seen in the previous DR challenges. To address all these limitations, we organized ''Diabetic Retinopathy—Grading and Image Quality Estimation Challenge'' (DeepDRiD) in ISBI 2020, and we designed three sub-challenges: (1) regular fundus DR grading for images in different quality, (2) image quality assessment for availability, and (3) ultra-widefield (UWF) DR grading for different device transferring. The following is the setup of our challenge:

- In the DeepDRiD, we presented regular fundus photographs for left and right eyes from each patient in dual views (macula-centered and optic-disc-centered) for the system development by participants.
- We provided a detailed quality assessment score for each image from the dataset.[11] We also provided a sub-challenge to assess image quality in four aspects: artifact, clarity, field definition, and overall score.
- We prepared a dataset from UWF fundus photography, containing one double-view shot per patient. This dataset offered the possibility to develop, validate, and test DR screening systems with multiple devices.

In this paper, we discuss details of the three best algorithms in each sub-challenge. All participants used convolutional

Table 1. International Clinical DR Severity Scale

Disease severity level	Descriptions	Findings observable on dilated ophthalmoscopy
Grade 0	no apparent retinopathy	no abnormalities
Grade 1	mild NPDR	microaneurysms only
Grade 2	moderate NPDR	between just microaneurysms and severe NPDR
Grade 3	severe NPDR	any of the following: more than 20 intraretinal hemorrhages in each of 4 quadrants; definite venous beading in more than 2 quadrants; prominent intraretinal microvascular abnormalities in more than 1 quadrant; no signs of PDR retinopathy
Grade 4	PDR	one or more of the following: neovascularization; vitreous/preretinal hemorrhage

PDR, proliferative diabetic retinopathy; NPDR, non-proliferative diabetic retinopathy.

Table 2. Image quality scoring criteria

Type	Image quality specification	Score
Artifact	no artifacts	0
	artifacts are outside the aortic arch with scope less than ¼ of the image	1
	artifacts do not affect the macular area with range less than ¼	4
	artifacts cover more than ¼ but less than ½ of the image	6
	artifacts cover more than ½ without fully covering the posterior pole	8
	cover the entire posterior pole	10
Clarity	clarity only level I vascular arch is visible	1
	level II vascular arch and a small number of lesions are visible	4
	level III vascular arch and some lesions are visible	6
	level III vascular arch and most lesions are visible	8
	level III vascular arch and all lesions are visible	10
Field definition	field definition do not include the optic disc and macula	1
	only contain either optic disc or macula	4
	contain optic disc and macula	6
	the optic disc or macula is outside the 1 papillary diameter and within the 2 papillary diameter range of the center	8
	the optic disc and macula are within 1 papillary diameter of the center	10
Overall quality	quality is not good enough for the diagnosis of retinal diseases	0
	quality is good enough for the diagnosis of retinal diseases	1

neural networks. Their algorithms differed mainly in terms of the detailed neural network architecture, training strategy, and pre- and post-processing methods. By examining their models, we validate the performance of image quality and DR grading, and we also summarize the relationship between these two tasks.

Methods

Materials

In the DeepDRiD challenge, we included patients from different projects in Shanghai, including participants in the Shanghai Diabetic Complication Screening Project, Nicheng Diabetes Screening Project, and Nation-wide Screening for Complications of Diabetes, for regular fundus images. The other part of fundus images included regular fundus images and UWF retinal images by retinal specialists at the outpatient ophthalmology clinic in the Sixth People's Hospital of Shanghai Jiao Tong University in China. From thousands of examinations available, we randomly selected 2,000 regular fundus images from 500 patients to form the regular fundus dataset. Each patient in the dataset has four fundus images, and each eye has two records, centered on the macula and optic disc. An example patient is shown in Figure S2A. Furthermore, 256 UWF images from another 128 patients formed our UWF dataset. The study was approved by the Ethics Committee of Shanghai Sixth People's Hospital and conducted in accordance with the Declaration of Helsinki. Informed consent was obtained from participants. The study was registered on the Chinese Clinical Trials Registry (ChiCTR.org.cn) under the identifier ChiCTR2000031184.

In addition to constructing the DeepDRiD dataset, we performed the following procedures to ensure image quality and accuracy of lesion diagnostic labels. Original retinal images were uploaded to the online platform, and the images of each eye were assigned separately to two authorized ophthalmologists. They labeled the images using an online reading platform and gave the image quality assessment scores and graded diagnosis of DR. The third ophthalmologist who served as the senior supervisor confirmed or corrected when the diagnostic results were contradictory. The final grading result was dependent on the consistency within these three ophthalmologists. Clinically, five levels of DR are distinguished, based on the International Clinical DR (ICDR)[4] classification scale: (1) no apparent retinopathy (grade 0), (2) mild NPDR (grade 1), (3) moderate NPDR (grade 2), (4) severe NPDR (grade 3), and (5) PDR (grade 4). Furthermore, the major factors affecting fundus image quality assessment are image artifact, low clarity, and low field definition, as shown in Figure S2B. The specific criteria of image quality assessment and DR grading can be seen in Tables 1 and 2. The DeepDRiD dataset is available to the public (Mendeley Data: https://doi.org/10.5281/zenodo.6452623).

For regular fundus images, the data were split into 60% for training (Regular Set-A: 300 patients, 1,200 images), 20% for testing (Regular Set-B: 100 patients, 400 images), and 20% for testing (Regular Set-C: 100 patients, 400 images). The UWF data were divided into UWF Set-A (77 patients, 154 images), UWF Set-B (25 patients, 50 images), and UWF Set-C (26 patients, 52 images). Moreover, Set-A and Set-B of regular fundus

Table 3. Basic characteristics of the patients in DeepDRiD dataset (mean ± SD)

	Regular fundus			UWF fundus		
DR levels	Set-A	Set-B	Set-C	Set-A	Set-B	Set-C
No. of images	1,200	400	400	77	25	26
No. of participants	300	100	100	154	50	52
Male (%)	51.00	44.00	46.00	54.55	57.69	48.00
Age (years)	70.63 ± 7.70	65.13 ± 1.89	61.36 ± 7.23	74.64 ± 4.86	64.96 ± 1.71	58.28 ± 4.88
BMI (kg m^{-2})	25.17 ± 3.13	24.88 ± 3.21	25.01 ± 2.58	24.90 ± 2.89	25.19 ± 2.61	24.06 ± 3.30
Waist (cm)	90.15 ± 9.24	88.36 ± 9.75	88.03 ± 8.87	88.43 ± 9.07	92.00 ± 8.86	84.73 ± 7.55

images and UWF images were provided to participants in model development, and the Set-C was used as an online validation set to evaluate the final performance. In addition, we provided patient-level DR grading results, including a comprehensive assessment of the DR grading results of both eyes. The distribution of DR severity in regular fundus images dataset (Regular Set-A, Regular Set-B, and Regular Set-C) is shown in Table 3. In the 256-image UWF fundus dataset, we collected labeling of DR grading levels according to the ICDR classification scale. The DR grading procedure for ophthalmologists is the same as that of the regular fundus. In this dataset, we only obtained two UWF fundus images from the right and left eye of each patient. We provided DR grading levels for the fundus image of each eye in the UWF image centered on the optic disc. Example figures of UWF fundus in different DR levels are shown in Figure S3. A detailed information of DeepDRiD dataset can be seen in Note S1.

Challenge setup

The DeepDRiD was composed of various stages, giving a well-organized work process to facilitate the success of contests. Figure 1 depicts the workflow of the overall organization of the challenge. The challenge was officially announced on the ISBI 2020 website on October 25, 2019. Following the DR challenge held with ISBI in 2018,[20] we decided to promote the progress further through the second challenge using a new dataset (DeepDRiD). The challenge was subdivided into three tasks as follows:

- Sub-challenge 1: DR disease grading: classification of fundus images according to the severity level of diabetic retinopathy using dual-view retinal fundus images.
- Sub-challenge 2: image quality estimation: fundus quality assessment for overall image quality, artifacts, clarity, and field definition.
- Sub-challenge 3: UWF fundus DR grading: explore the generalizability of a DR grading system. The robust and generalizable models were expected to be developed to solve practical clinical issues.

We set up a website to share information about the challenge and provide an interface for all challenge-related issues. The challenge website is accessible directly at https://isbi.deepdr.org. On the website, the participants could register and find a general overview of the challenge, including the deadlines, a brief description of the biomedical background of the problem, a description of the dataset, the rules of the challenge, the evaluation metrics, and Python code snippets for accessing the images and the annotations. Finally, the participants could submit their results and access a forum to ask questions and provide comments through the website. It consisted of an open-testing round (Regular Set-B and UWF Set-B) for teams to refine and calibrate their models, and a final evaluation round (Regular Set-C and UWF Set-C). Participants were granted access to the dataset, forum, and submission system after they registered and accepted the rules of the challenge. Anonymous participation was not allowed. The complete DeepDRiD datasets were shared on GitHub (Mendeley Data: https://doi.org/10.5281/zenodo.6452623). The challenge aimed for a fair comparison of algorithms. Due to the large size of the public dataset in fundus images, participants were allowed to use other data sources but were required to mention which data they used.

The participants had to submit their results as CSV files through the challenge website. The deadline for submissions was March 4, 2020. A maximum of three submissions was allowed per participant, with a four-page ISBI style paper accompanying each submission describing their methods. The three submissions had to be methodologically different. Resubmissions with simple hyper-parameter tuning were not allowed. During the workshop at ISBI 2020, we presented the challenge results and invited the top three teams to present their methods. The results, presentations, and algorithms of participants were shared on the challenge website after the workshop. Subsequently, the challenge was reopened for registration and submissions. In submission result analysis, we used quadratic weighted kappa (κ_ω) as the assessment metric for sub-challenges. Moreover, in sub-challenge 2, the overall quality is evaluated by accuracy. The details of the evaluation method can be seen in supplemental information: evaluation metrics.

RESULTS

We had 574 registered participants before March 1, 2019, when the test dataset was released. The teams explored a wide range of machine learning and deep learning models, ranging from CatBoost,[27] LightGBM,[28] XGboost,[29] VGG,[30] ResNet,[31] SE-ResNeXt,[32] to EfficientNet,[33] and combinations of several types of models. In total, 34 teams submitted their models in our challenge. To help the participating teams avoid overfitting problems, we also provided a separate validation set (Regular Set-B and UWF Set-B) during the competition to help them validate the model results. In Figure 2, we gave the results of three sub-challenges in the rank scores.

Summary of competing solutions

We only present the methodology and results of the top three best-performing algorithms in each sub-challenge to keep the

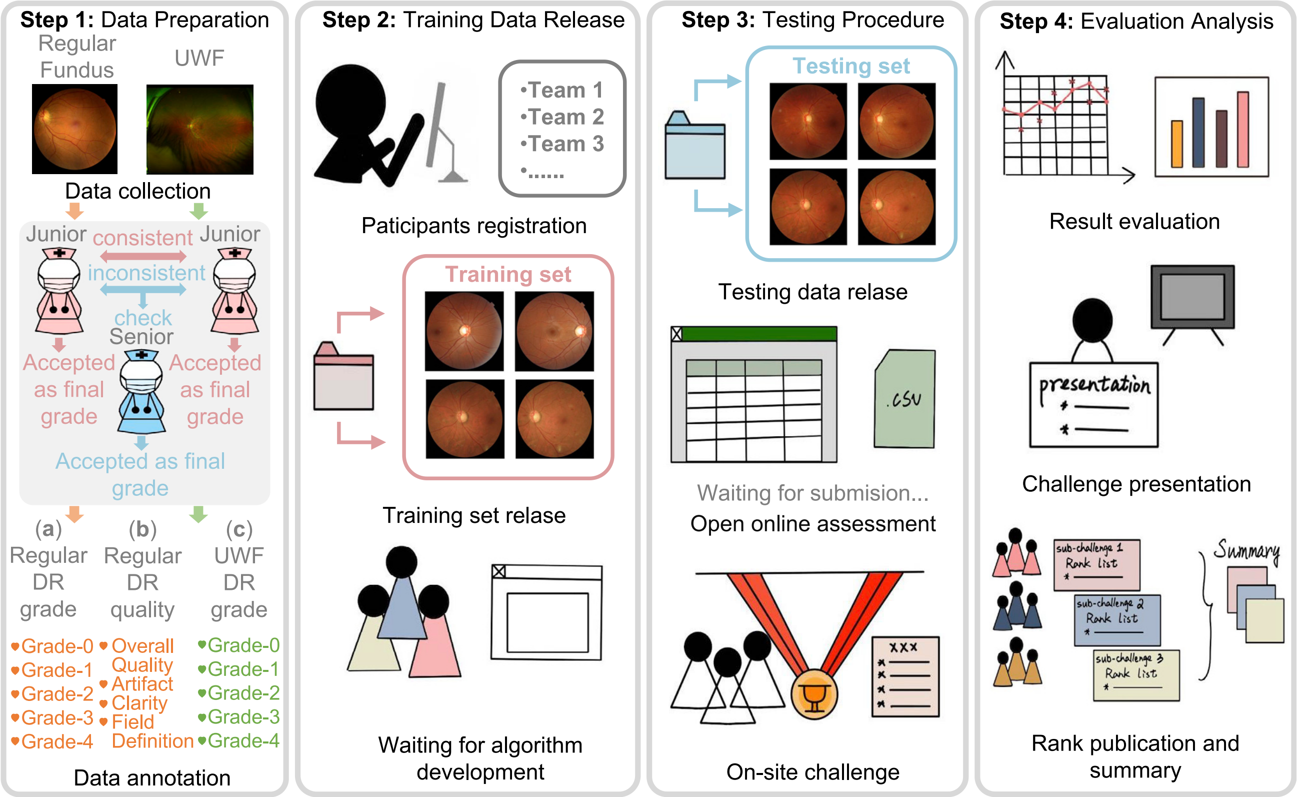


Figure 1. Workflow of the ISBI 2020: Diabetic Retinopathy—Grading and Image Quality Estimation Challenge

paper concise. We discuss the algorithms of the nine teams in terms of the following steps: data preprocessing, data augmentation, model pre-training, and training strategies of classifying deep learning models. We provide a summary and then discuss each of these four steps. The detail models and specific training strategies are shown in Note S2. Moreover, we summarize the commonalities in the good results achieved by these team approaches. These teams all considered the background of medical expertise and considered the diagnostic processes of professional physicians in the preprocessing of the data, the training of the models, and the integration of the final results.

- The improvement of model generalization performance is achieved by pre-training the model with extensive use of routine fundus images published in publicly available datasets and DR grading results from professional physicians.
- Considering the task-to-task correlation, knowledge migration from source to target data is utilized, enabling the model to learn important information quickly.
- Simultaneous training and integration of multiple models are used to improve the performance and performance of the models using training strategies in the field of deep learning.

The winning teams in three sub-challenges have different characteristics. In sub-challenge 1 (DR grading using regular fundus), the winning team did not use complex data preprocessing and augmentation operations, but used advanced deep learning training tools from the training means. In sub-challenge 2 (image quality assessment), the winning team used rich data preprocessing and augmentation operations to design the model. The winning team in sub-challenge 3 (UWF DR grading) won the competition by pre-training and knowledge transfer of large-scale data.

Data preprocessing

We analyzed different preprocessing steps used in each of three sub-challenges: DR grading based on regular fundus; image quality assessment based on regular fundus; and DR grading based on UWF fundus. In DR severity grading of regular fundus images, public dataset providing large size of regular fundus images and their DR grading results, such as IDRiD,[20] Kaggle 2015,[21] Messidor,[22] Kaggle 2009,[23] ROC,[24] E-Ophtha,[25] DiaretDB,[26] and REFUGE 2,[34] were used. In a previous study, general preprocessing methods were introduced to improve model performance of DR grading. Some teams adopted these preprocessing algorithms, including Ben's preprocessing method,[35] image transformation based on bilinear interpolation, reducing the black edges of fundus images, and so on. In the image quality assessment task based on the regular fundus, the preprocessing algorithms used by participants were fundamentally the same as the DR severity grading task also based on the regular fundus. Moreover, due to difference between the regular fundus and UWF fundus, the preprocessing steps were different in the DR grading task based on regular fundus and UWF fundus. In the UWF fundus, all teams used the center-cut method to cut the edge of the UWF fundus images. For more details, we refer to Table 4.

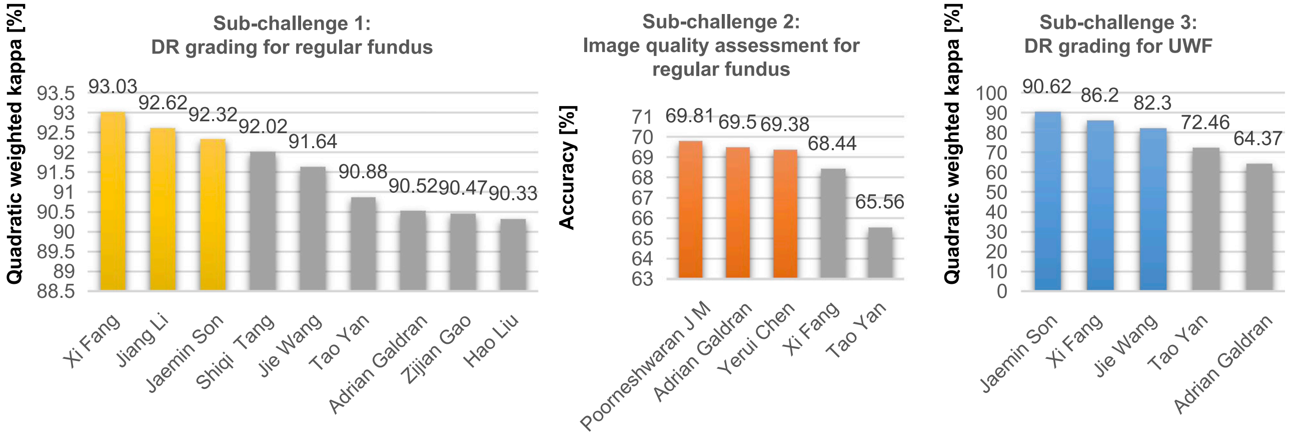


Figure 2. Bar chart for leaderboard in three sub-challenges

The colored bars indicate the top three teams in each challenge.

Data augmentation

The color distribution of fundus images can influence the robustness of the convolution neural networks (CNNs). Most teams used data augmentation methods, such as color adjustment, mirroring, rotation, and so on, to maintain the generality of CNNs. In sub-challenge 1, two teams adopted the same mirroring method: horizontal flip, vertical flip, and horizontal and vertical flip; they also used rotation augmentation with different rotation angles. As most teams participating in sub-challenges 2 and 3 used the pre-trained network migration based on sub-challenge 1, they did not use data augmentation methods. In sub-challenge 2, only the top 1 team used additional color adjustment methods, i.e., CutMix,[36] RICAP,[37] and Mixup.[38] Furthermore, the top 1 team in sub-challenge 3 adopted plentiful augmentation strategies. The results from sub-challenges 2 and 3 show that, although pre-trained network transfer helps the network learn new tasks quickly, the use of data augmentation methods is still helpful in improving network results. Table 5 shows the detailed data augmentation method adopted by nine teams.

Model pre-training

Many eye diseases are diagnosed based on fundus images. Thus, it is common for public datasets in fundus images to build models for different eye diseases. In our challenge, most teams selected to use model pre-training to improve their model ability. Table 6 shows model pre-training details.

Classifying deep learning models

In three sub-challenges, all teams used current deep learning models to construct classification frameworks. Most teams adopted EfficientNet[33] as their deep learning backbones, and obtained great model performance, whereas some teams selected classical ResNet[31] and its variant SE-ResNeXt.[32] A team in sub-challenge 2 used a private dataset with regular fundus image and pixel-level structure labels, and selected UNet[39] and VGG[30] as their deep learning classification model. Most teams chose regular classification loss functions, such as cross-entropy loss (CE), L1 loss, and smooth L1 loss. One team in sub-challenge 2 proposed and adopted cost-sensitive loss.[40] The teams that selected different training strategies to develop deep learning models are detailed in Table 7.

Solution results

To fairly evaluate the performance of the individual competition team models, the quadratic weighted kappa score κ_ω was used to rank the algorithms. The κ_ω ranged from 0.9303 to 0.9033 for all nine participating teams in sub-challenge 1, from 0.6981 to 0.6938 for the sub-challenge 2, and from 0.9062 to 0.6437 for sub-challenge 3. In sub-challenge 1, almost all teams achieved good performance (>0.90); in sub-challenge 2, almost all teams achieved unsatisfactory performance (<0.70). This may be partly due to the fact that the teams in the competition did not take into account well the unevenness of the categories, and the relatively small differences between classes that are difficult to extract. In sub-challenge 3, all the teams performed evenly in distribution (0.60–0.90). Correlation between different fundus images was considered, and better accuracy was achieved using a team of transfer learning and sliding window learning. For the scores of the top 3 teams in each sub-challenge, we refer to Table 8. We also give a summary of the participation of these nine teams for all sub-challenges in Table S4. The performances of the proposed methods on the final validation (Set-C) are shown in six subtasks (divided into three sub-challenges). The leaderboard ranks in the three sub-challenges are also illustrated.

Sub-challenge 1: DR disease grading

This section presents the performance of all competing solutions in the DR grading task using regular fundus pictures. The results received from the participating teams were analyzed using κ_ω as a validation measure. κ_ω was calculated on the validation set (Regular Set-C) for each of the different techniques. Of the 34 participating teams in the challenge, 11 teams participated in sub-challenge 1. Of these 11 teams, 9 (see Table S5) performed well in the DR grading task and then were invited to participate in the challenge workshop. The top three groups were those of Xi Fang, Jiang Li, and Jaemin Son. The classification results of the three teams reflect that all of their models achieved good classification performance, with sensitivity and specificity comparable with physicians on the grading from normal to PDR. In addition, the classification results showed a slightly higher degree of confusion for mild lesions than for moderate and severe lesions. In Note S3, we detail the model performance and result analysis.

Table 4. Differences in preprocessing

RK	Cut	Color	Resize	Filling
Sub-challenge 1: DR grading				
1	N	N	N	N
2	black edge	Ben's[35]	Bi (512)	N
3	black edge	N	Bi (1,024)	N
Sub-challenge 2: image quality assessment				
1	N	N	Bi (512)	N
2	N	N	N	N
3	black edge	N	N	flip
Sub-challenge 3: DR grading based on UWF fundus				
1	center	N	N	N
2	center	N	N	N
3	N	N	N	N

Black edge, cut the black edges in the fundus; center, preserve the center of the image as input; Ben's, Ben's preprocessing algorithm;[35] Bi(*i*), use bilinear interpolation to resize the fundus image to *i* pixels size; flip, use a symmetrical flip pattern to fill the black edges; N, never use this strategy; RK, rank.

Table 5. Differences in data augmentation

RK	Mirroring	Rotation	Color	Other
Sub-challenge 1: DR grading				
1	N	N	N	N
2	H/V/HV	R: −30, +30	N	N
3	H/V/HV	R: −20, +20	ID/N	R/ET/GT/AT
Sub-challenge 2: image quality assessment				
1	N	N	CM/RC/MU	N
2	N	N	N	N
3	N	N	N	N
Sub-challenge 3: DR grading based on UWF fundus				
1	H/V/HV	R: −20, +20	ID/N	R/ET/GT/AT
2	N	N	N	N
3	N	N	N	RCC

H, horizontal flip; V, vertical flip; HV, horizontal and vertical flip; R, min degree, max degree:rotation angle; ID, image disturbance; N, noise; R, resize; ET, elastic transformation; GT, grid transformation; AT, affine transformation; RCC, random center cut; CM,[36] RC,[37] and MU,[38] preprocessing method in reference; RK, rank.

Sub-challenge 2: Image quality estimation

This task was performed using the validation algorithm described in Note S2 on Set-C to evaluate four aspects of image quality: artifacts, clarity, field definition, and overall quality. The algorithm produced scores for the above four aspects. The best-performing solution in the on-site sub-challenge two was proposed by Poorneshwaran J M, followed by Adrian Galdran and Yerui Chen. For the teams performing poorly in several tasks of image quality detection, their overall accuracy of image quality detection was also not high. The main reason seems to be the inaccurate differentiation of the degree of DR image quality due to the uneven distribution of classes in the dataset and the relatively high degree of similarity between classes. The detailed result can be seen in Note S3.

Sub-challenge 3: UWF fundus DR grading

The results for DR grading of UWF fundus images were obtained by the same evaluation method as used for sub-challenge 1 using κ_ω. Table S6 shows the results of the field evaluation, summarizing the performance of all participating algorithms in the UWF fundus DR grading task. Jaemin Son developed the winning method for the UWF fundus DR grading, and Jaemin Son, Xi Fang, and Jie Wang won the best top 3 performers in this task.

DISCUSSIONS

Summary of holding and analyzing the challenge

In this paper, we present the details of the DeepDRiD challenge, including relevant information regarding the dataset, evaluation metrics for multiple sub-challenges of the competition, the organization of the challenge, solutions, and results by the participating teams on all sub-challenges. The sub-challenges included grading DR severity, quality detection and assessment of fundus photo images, and UWF fundus images DR grading. With 34 teams participating the challenge and reporting the results, we consider our challenge successful. We did our best to create a relevant, stimulating, and fair competition for advancing the collective knowledge of the research community.

The best methods for DR lesion severity grading used a considerable number of common tips: (1) efficient extraction of features through data augmentation, (2) transfer learning of large amounts of fundus data with and without physician labels, and (3) loss function modification. In addition, many grading networks used the EfficientNet-based framework[33] to learn grading features quickly and efficiently, which improved the performance of the models. The rich parameter adjustment methods and model fusion methods also provided new ideas to further solve the DR grading problem. In the quality assessment task, the accuracy of image quality detection ranged between 0.68 and 0.70. The results did not reach the performance required for clinically feasible automatic screening of good quality fundus images; therefore, there is still much work to do in image quality assessment. Attention must be paid to features of both artifacts and clarity to improve the overall assessment results considering the misclassification cases. In sub-challenge 3, the results of five teams were used for evaluation. We observe that using those readily available regular fundus images for knowledge transfer has a very significant effect on the DR grading task for the same UWF images of the fundus of the eye.

Limitations of the study

This challenge provided data collected in routine clinical practice using an acquisition protocol consistent with all images. The data were acquired with the same camera simultaneously after pupil dilation and followed to provide annotations corresponding to the quality assessment protocol. Several experts jointly evaluated the images in this dataset, and images disagreed by experts were excluded from the dataset. Even after these efforts (for providing the best possible data), the annotation process (especially for image quality) remained inherently subjective.

Table 6. Differences in model pre-training

RK	Pre-training dataset
Sub-challenge 1: DR grading	
1	Kaggle2015 + APTOS
2	Kaggle2015 + APTOS
3	labeled and unlabeled dataset
Sub-challenge 2: image quality assessment	
1	ImageNet
2	Kaggle2015
3	private fundus lesion segmentation data
Sub-challenge 3: DR grading based on UWF fundus	
1	labeled and unlabeled
2	Kaggle2015 + AOTOS
3	N

The public datasets used are Kaggle2015,[21] APTOS.[47] Labeled: Kaggle2015,[21] APTOS,[47] and IDRiD;[20] unlabeled: REFUGE,[34] MESSIOR,[22] and E-ophtha.[25] RK, rank.

Thus, manual judgment is a limiting factor in the method development, especially for the methods trained and evaluated in a supervised manner. Our challenge provides the potential to develop DR lesion grading solutions, fundus image quality assessment, and DR grading using UWF fundus images. Despite the complexity of the tasks and also just 1.5 months for method development, it still received a very positive response from the community. Nevertheless, there is still room for improvement, especially in evaluation of image quality. Therefore, although the competition is over, the dataset is still publicly available for research purposes, to attract more researchers to study the problem and develop new solutions to meet current and future clinical standards.

Insights on the future directions

Based on our analysis of the organization of this challenge and the results from the challenge, we propose the following ideas for future directions. First, almost all teams in this challenge used deep learning models as the main network framework to solve this problem. The results also show that the deep learning models do achieve good results, which demonstrates the great potential of deep models in this problem. Second, in the pre-training of the model, almost all teams used a wide range of general fundus images for model pre-training and parameter migration. This is based on the relatively extensive research interest and a large number of datasets publicly available for this problem on the one hand, and the significant importance of pre-training for model performance improvement on the other hand. Finally, in different subtasks, we can find that the models that achieved victory have different characteristics, some are more focused on preprocessing and augmentation methods and some are more focused on the model architecture and training means. This means that developing models for specific medical problems requires more problem-specific analysis.

Suggestions for organizing medical grand challenges

To help the research community better organize medical grand challenges, we also give a few of our tests. First, the motivation for the challenge needs to come from the clinician's real-world problems. For example, in our challenge, all three subtasks come from the difficulties and challenges encountered in automated deep learning screening during DR screening. In addition, reasonable and compliant access to data prior to organizing the challenge requires that we communicate and collaborate with clinicians as early as possible. Second, the organization, promotion, and conduct of the challenge needed to be as rich in diversity as possible: diversity of competition organizers, diversity of participants, etc. (from different countries and regions, different professional backgrounds, etc.). Finally, a long research base will also help the organizers to better organize the competition and sustainably lead the direction.

Table 7. Differences in deep learning models

RK	Model frameworks	Loss function	Training strategies
Sub-challenge 1: DR grading			
1	EfficientNet[33]	SL1	MMoE + GMP + ES + OHEM + CV + O + T
2	EfficientNet[33]	SL1 + CE + DV + PL	CV + TTA
3	EfficientNet[33]	L1 + CE(5 class)	PLT
Sub-challenge 2: image quality assessment			
1	SE-ResNeXt[32]	CE	TL
2	ResNet[31]	CS + L1	TL
3	VGG,[30] UNet[39]	CE	TL
Sub-challenge 3: DR grading based on UWF fundus			
1	EfficientNet[33]	L1 + CE(5 class)	PLT
2	EfficientNet[33]	SL1	MMoE + GMP + ES + OHEM + CV + O + T
3	EfficientNet[33]	CE	TL

SL1, smooth L1 loss; CE, cross-entropy loss; DV, dual view loss; PL, patient-level loss; CS, cost-sensitive loss;[40] L1, L1 loss; CE(5 class), mean loss of 5 class (one versus others); MMoE, multi-gate mixture of expert;[41] GMP, generalized mean pooling;[42] OHEM, online hard example mining;[43,44] CV, cross-validation; O, oversampling; ES, early stopping; TL, transfer learning; TTA, test time augmentation;[45,46] PLT, pseudo-labeled and labeled training.

Conclusion

By leveraging hospital research data and physician resources, we provide a finely labeled dataset of realistic DR screening scenarios that demonstrate the diagnostic potential of the DeepDRiD challenge models on conventional DR grading, DR image quality assessment, and ultra-wide angle fundus DR grading. These models obtained comparable diagnostic performance with general ophthalmologists on DR grading and preliminary attempts on image quality assessment. Furthermore, these new deep learning prediction models and their training strategies can be used to enhance the diagnostic capabilities of healthcare workers to improve the accuracy of DR screening in true screening scenarios. Nevertheless, there is still a clear opportunity to further improve the models in this competition. We believe that, with access to higher quality and more comprehensive image quality assessment data, as well as a wider range of challenge participants, more accurate models could be developed.

Table 8. DeepDRiD online leaderboard

Rank	Team	Affiliation	Score
Sub-challenge 1			
1	Xi Fang et al.	Shanghai Jiao Tong University	0.9303
2	Jiang Li et al.	Shanghai Jiao Tong University	0.9262
3	Jaemin Son et al.	VUNO Inc.	0.9232
Sub-challenge 2			
1	Poorneshwaran J M et al.	Healthcare Technology Innovation Center	0.6981
2	Adrian Galdran et al.	Bournemouth University	0.6950
3	Yerui Chen et al.	Nanjing University of Science and Technology	0.6938
Sub-challenge 3			
1	Jaemin Son et al.	VUNO Inc.	0.9062
2	Xi Fang et al.	Shanghai Jiao Tong University	0.8620
3	Jie Wang et al.	Beihang University	0.8230

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Bin Sheng (shengbin@sjtu.edu.cn).

Materials availability

This study did not generate any new materials.

Data and code availability

The DeepDRiD dataset is available at https://github.com/deepdrdoc/DeepDRiD (Mendeley Data: https://doi.org/10.5281/zenodo.6452623).

Evaluation metrics

For DR disease-grading tasks, in sub-challenges 1 and 3, the quadratic weighted kappa (κ_ω) was used as the evaluation metric to determine the performance of the participating algorithms. Submissions were scored based on the quadratic weighted kappa, κ_ω, which measures the agreement between two ratings (ground-truths results and submitted results). This metric varied from 0 (random agreement between raters) to 1 (complete agreement between raters). If there was less agreement between the raters than expected by chance, the metric could go below 0. The quadratic weighted kappa, κ_ω, was calculated between the scores, which were expected/known, and the predicted scores.

The results had five possible ratings: 0, 1, 2, 3, and 4. The quadratic weighted kappa was calculated as follows. First, an $N \times N$ histogram matrix, O, was constructed, such that it corresponded to the number of adoption records that had a rating of i (actual) and received a predicted rating, j. An $N \times N$ matrix of weights, w, was calculated based on the difference between the actual and predicted rating scores. An $N \times N$ histogram matrix of expected ratings, E, was calculated, assuming no correlation between rating scores. This was calculated as the outer product between the actual rating's histogram vector of ratings and the predicted rating's histogram vector of ratings, normalized such that E and O had the same sum. From these three matrices, the quadratic weighted kappa was calculated. The κ_ω metric is expressed as

$$\kappa_w = 1 - \frac{\sum_{i,j} w_{i,j} \cdot O_{i,j}}{\sum_{i,j} w_{i,j} \cdot E_{i,j}}. \qquad \text{(Equation 1)}$$

The weight penalization, $w_{i,j}$, is defined by $w_{i,j} = \frac{(i-j)^n}{(C-1)^n}$, where C is the number of classes. The values of $n = 1$ and $n = 2$ lead to linear and quadratic penalizations, respectively. The values of κ_ω is in the interval of $\kappa_w \in [-1, 1]$, where -1 means perfect symmetric disagreement and 1 means perfect agreement.

In sub-challenge 2, the scoring metric was classification accuracy, as described as

$$Accuracy = \frac{TP + TN}{N}, \qquad \text{(Equation 2)}$$

where TP is true positive samples, FP is false positive samples, and $N = TP + FP + TN + FN$ is the total numbers.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2022.100512.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (61872241, 81870598, and 82022012); the Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102); the College-level Project Fund of Shanghai Jiao Tong University Affiliated Sixth People's Hospital (ynlc201723 and ynlc201909); and the Medical-industrial Cross-fund of Shanghai Jiao Tong University (YG2022QN089).

AUTHOR CONTRIBUTIONS

Conceptualization, R.L., B.S., and P.Z.; resources, X.W., Q.W., H. Li, and W.J.; methodology, R.L., X.W., L.D., X.F., T.Y., J.S., S.T., J.L., Z.G., A.G., P.J.M., H. Liu, J.W., and Y.C.; software, C.D., H.S., and M.C.; formal analysis, X.W., Q.W., H. Li, P.P., G.S.W.T., X.Y., D.S., and W.J.; writing – original draft, R.L. and P.Z.; writing – review & editing, X.W., Q.W., B.S., H. Li, P.Z., and W.J.; supervision, B.S., H. Li, P.Z., D.S., and W.J.; funding acquisition, X.W., H. Li, and B.S.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We worked to ensure gender balance in the recruitment of human subjects. We worked to ensure ethnic or other types of diversity in the recruitment of human subjects. The author list of this paper includes contributors from the location where the research was conducted who participated in the data collection, design, analysis, organization, and participation of the challenge.

Received: March 14, 2022
Revised: March 28, 2022
Accepted: April 25, 2022
Published: May 20, 2022

REFERENCES

1. Atlas, I.D.F.D. (2017). Brussels, belgium: International Diabetes federation (Int. Diabet. Federat. (IDF)).
2. Hutchinson, A., McIntosh, A., Peters, J., O'keeffe, C., Khunti, K., Baker, R., and Booth, A. (2000). Effectiveness of screening and monitoring tests for diabetic retinopathy : a systematic review. Diabet. Med. *17*, 495–506.
3. Reichel, E., and Salz, D. (2015). Diabetic retinopathy screening. Managing Diabetic Eye Disease in Clinical Practice, pp. 25–38.
4. Organization, W.H. (2005). Prevention of blindness from diabetes mellitus. In Report of a WHO consultation (Geneva, Switzerland: WHO), pp. 1–48.
5. Wei, W., and ACY, L. (2018). Diabetic retinopathy: pathophysiology and treatment. Int. J. Mol. Sci. *19*, 1816.
6. Ruta, L.M., Magliano, D.J., Lemesurier, R., Taylor, H.R., Zimmet, P.Z., and Shaw, J.E. (2013). Prevalence of diabetic retinopathy in type 2 diabetes in developing and developed countries. Diabet. Med. *30*, 387–398.
7. Kung, K., Chow, K.M., Hui, E.M.T., Leung, M., Leung, S.Y., Szeto, C.C., Lam, A., and Li, P.K.T. (2014). Prevalence of complications among Chinese diabetic patients in urban primary care clinics: a cross-sectional study. BMC Prim. Care *15*, 8.

8. Hu, Y., Teng, W., Liu, L., Chen, K., Liu, L., Hua, R., Chen, J., Zhou, Y., and Chen, L. (2015). Prevalence and risk factors of diabetes and diabetic retinopathy in liaoning province, China: a population-based cross-sectional study. PLoS One *10*, e0121477.

9. Pang, C., Jia, L., Jiang, S., Liu, W., Hou, X., Zuo, Y., Gu, H., Bao, Y., Wu, Q., Xiang, K., et al. (2012). Determination of diabetic retinopathy prevalence and associated risk factors in Chinese diabetic and pre-diabetic subjects: Shanghai diabetic complications study. Diabetes Metab. Res. Rev. *28*, 276–283.

10. Lian, J.X., Gangwani, R.A., McGhee, S.M., Chan, C.K.W., Lam, C.L.K.; Primary Health Care Group, and Wong, D.S.H. (2016). Systematic screening for diabetic retinopathy (dr) in Hong Kong: prevalence of dr and visual impairment among diabetic population. Br. J. Ophthalmol. *100*, 151–155.

11. Dai, L., Wu, L., Li, H., Cai, C., Wu, Q., Kong, H., Liu, R., Wang, X., Hou, X., Liu, Y., et al. (2021). A deep learning system for detecting diabetic retinopathy across the disease spectrum. Nat. Commun. *12*, 3242.

12. Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA *316*, 2402–2410.

13. Ting, D.S.W., Cheung, C.Y.L., Lim, G., Tan, G.S.W., Quang, N.D., Gan, A., Hamzah, H., Garcia-Franco, R., San Yeo, I.Y., Lee, S.Y., et al. (2017). Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. Facial Plast. Surg. Aesthet. Med. *318*, 2211–2223.

14. van der Heijden, A.A., Abramoff, M.D., Verbraak, F., van Hecke, M.V., Liem, A., and Nijpels, G. (2018). Validation of automated screening for referable diabetic retinopathy with the idx-dr device in the hoorn diabetes care system. Acta Ophthalmol. *96*, 63–68.

15. Li, Z., Keel, S., Liu, C., He, Y., Meng, W., Scheetz, J., Lee, P.Y., Shaw, J., Ting, D., Wong, T.Y., et al. (2018). An automated grading system for detection of vision-threatening referable diabetic retinopathy on the basis of color fundus photographs. Diabetes Care *41*, 2509–2516.

16. Liu, Y., Wang, M., Morris, A.D., Doney, A.S.F., Leese, G.P., Pearson, E.R., and Palmer, C.N.A. (2013). Glycemic exposure and blood pressure influencing progression and remission of diabetic retinopathy: a longitudinal cohort study in godarts. Diabetes Care *36*, 3979–3984.

17. Araújo, T., Aresta, G., Mendonça, L., Penas, S., Maia, C., Carneiro, Â., Mendonça, A.M., and Campilho, A. (2020). DR | GRADUATE: uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. Med. Image Anal. *63*, 101715.

18. He, A., Li, T., Li, N., Wang, K., and Fu, H. (2021). CABNet: category attention block for imbalanced diabetic retinopathy grading. IEEE Trans. Med. Imag. *40*, 143–153.

19. Zhou, Y., Jiang, Q., Ma, S., Zhou, X., and Shao, L. (2021). Effect of quercetin on the in vitro Tartary buckwheat starch digestibility. Int. J. Biol. Macromol. *183*, 818–830.

20. Porwal, P., Pachade, S., Kokare, M., Deshmukh, G., Son, J., Bae, W., Liu, L., Wang, J., Liu, X., Gao, L., et al. (2020). Idrid: diabetic retinopathy - segmentation and grading challenge. Med. Image Anal. *59*, 101561.

21. EyePACS. (2015). Diabetic retinopathy detection. Available. https://www.kaggle.com/c/diabetic-retinopathy-detection/. July 28th, 2015.

22. Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., Gain, P., Ordonez, R., Massin, P., Erginay, A., and Charton, B. (2014). Feedback *on a p*ublicly distributed image database: the messidor database. Image Anal. Stereol. *33*, 231–234.

23. Cuadros, J., and Bresnick, G. (2009). EyePACS: an adaptable telemedicine system for diabetic retinopathy screening. J. Diabetes Sci. Technol. *3*, 509–516.

24. Niemeijer, M., van Ginneken, B., Cree, M.J., Mizutani, A., Quellec, G., Sanchez, C.I., Zhang, B., Hornero, R., Lamard, M., Muramatsu, C., et al. (2010). Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs. IEEE Trans. Med. Imag. *29*, 185–195.

25. Decencière, E., Cazuguel, G., Zhang, X., Thibault, G., Klein, J.C., Meyer, F., Marcotegui, B., Quellec, G., Lamard, M., Danno, R., et al. (2013). Teleophta*: mach*ine learning and image processing methods for teleophthalmology. Irbm *34*, 196–203.

26. Kauppi, T., Kamarainen, J.K., Lensu, L., Kalesnykiene, V., Sorri, I., Uusitalo, H., and Kälviäinen, H. (2012). A framework for constructing benchmark databases and protocols for retinopathy in medical image analysis. In IScIDE'12 Proceedings of the third Sino-foreign-interchange conference on Intell. Sci. Intell. Data Eng., pp. 832–843.

27. Dorogush, A.V., Ershov, V., and Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. Preprint at arXiv. https://doi.org/10.48550/arXiv.1810.11363.

28. Ke, G., et al. (2017). LightGBM: a highly efficient gradient boosting decision tree. In Proc. Neurips, pp. 3146–3154.

29. Chen, T., and Guestrin, C. (2016). XGBoost: a scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794.

30. Simonyan, K., and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. Preprint at arXiv. https://doi.org/10.48550/arXiv.1409.1556.

31. 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.

32. Hu, J., Shen, L., Albanie, S., Sun, G., and Wu, E. (2020). Squeeze-and-excitation networks. IEEE Trans. Pattern Anal. Mach. Intell. *42*, 2011–2023.

33. Tan, M., and Le, Q.V. (2019). EfficientNet: rethinking model scaling for convolutional neural networks. In Proc. ICML, pp. 6105–6114.

34. REFUGE2 (2020). Retinal Fundus Glaucoma Challenge Edition 2. Lima, Peru. https://refuge.grand-challenge.org/.

35. Graham, B. (2014). Spatially-sparse convolutional neural networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1409.6070.

36. Yun, S., Han, D., Chun, S., Oh, S.J., Yoo, Y., and Choe, J. (2019). CutMix: regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 6022–6031.

37. Takahashi, R., Matsubara, T., and Uehara, K. (2018). RICAP: random image cropping and patching data augmentation for deep cnns. In Proc. ACML, J. Zhu and I. Takeuchi, eds., pp. 786–798.

38. Zhang, H., Cissé, M., Dauphin, Y.N., and Lopez-Paz, D. (2018). mixup: beyond empirical risk minimization. In Proc. ICLR.

39. Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: convolutional networks for biomedical image segmentation. In Diabet. Foot. Ulcers. Grand. Chall. (2021), pp. 234–241.

40. Galdran, A., Dolz, J., Chakor, H., Lombaert, H., and Ben Ayed, I. (2020). Cost-sensitive regularization for diabetic retinopathy grading from eye fundus images. In Comput. Diffus. MRI. (2019), pp. 665–674.

41. Ma, J., Zhao, Z., Yi, X., Chen, J., Hong, L., and Chi, E.H. (2018). Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In Proc. KDD, pp. 1930–1939.

42. Radenovic, F., Tolias, G., and Chum, O. (2019). Fine-tuning CNN image retrieval with no human annotation. IEEE Trans. Pattern Anal. Mach. Intell. *41*, 1655–1668.

43. Schroff, F., Kalenichenko, D., and Philbin, J. (2015). FaceNet: a unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 815–823.

44. Shrivastava, A., Gupta, A., and Girshick, R.B. (2016). Training region-based object detectors with online hard example mining. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 761–769.

45. Bahat, Y., and Shakhnarovich, G. (2020). Classification confidence estimation with test-time data-augmentation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2006.16705.

46. Kandel, I., and Castelli, M. (2021). Improving convolutional neural networks performance for image classification using test time augmentation: a case study using MURA dataset. Health Inf. Sci. Syst. *9*, 33.

47. APTOS. (2019). The 4th asia pacific tele-ophthalmology society (aptos) symposium. Available: https://www.kaggle.com/c/aptos2019-blindness-detection