Title: IDRiD: Diabetic Retinopathy - Segmentation and Grading Challenge

Abstract: Diabetic Retinopathy (DR) is the most common cause of avoidable vision loss, predominantly affecting the working age population across the globe. Screening for DR, coupled with timely consultation and treatment, is a globally trusted policy to avoid vision loss. However, the implementation of DR screening programs is challenging due to the scarcity of medical professionals able to screen a growing global diabetic population at risk for DR. Computer-aided disease diagnosis in retinal image analysis could provide a sustainable approach for such large-scale screening effort. The recent scientific advances in computing capacity and machine learning approaches provide an avenue for biomedical scientists to reach this goal. Aiming to advance the state-of-the-art in automatic DR diagnosis, the Grand Challenge on "Diabetic Retinopathy - Segmentation and Grading" was organized in conjunction with the IEEE International Symposium on Biomedical Imaging (ISBI - 2018). In this paper, we report the set-up and results of this challenge that is primarily based on Indian Diabetic Retinopathy Image Dataset (IDRiD). There were three principal sub-challenges: lesion segmentation, disease severity grading, and localization of retinal landmarks and segmentation. These multiple tasks in this challenge allow to test the generalizability of the algorithms, and this is what makes it different from the existing ones. It received a positive response from a scientific community with 148 submissions from 495 registrations effectively entered in this
challenge. This paper outlines the challenge, its organization, the dataset used, evaluation methods and results of top performing participating solutions. We observe that the top performing approaches utilize a blend of clinical information, data augmentation, and the ensemble of models. These findings have the potential to enable new developments in retinal image analysis and image-based DR screening in particular.

Research Data Related to this Submission
-------------------------------------
Title: Indian Diabetic Retinopathy Image Dataset (IDRiD)
Repository: IEEE DataPort
https://ieee-dataport.org/open-access/indian-diabetic-retinopathy-image-dataset-idrid
IDRiD: Diabetic Retinopathy - Segmentation and Grading Challenge


*Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded, India
bSchool of Biomedical Informatics, The University of Texas Health Science Center at Houston, USA
cEye Clinic, Sushrusha Hospital, Nanded, Maharashtra, India
dVUNO Inc., Seoul, Republic of Korea
ePing An Technology (Shenzhen) Co., Ltd, China
fiFLYTEK Research, Hefei, China
gSchool of Electrical and Computer Engineering, University of Oklahoma, USA
hSamsung Advanced Institute for Health Sciences & Technology (SAHST), Sungkyunkwan University, Seoul, Republic of Korea
iDepartment of Computer Science, University of North Carolina at Charlotte, USA
jSchool of Information Science and Engineering, Shandong Normal University, China
kDepartment of Radiology and BRIC, the University of North Carolina at Chapel Hill, USA
lCleerly Inc., New York, United States
mVirginia Tech, Virginia, United States
nUniversity at Buffalo, New York, United States
oUniversity of Debrecen, Faculty of Informatics 4002 Debrecen, POB 400, Hungary
pIndividual Researcher, India
qCollege of Computer Science and Technology, Zhejiang University, Hangzhou, China
rMachine Learning For Bioimage Analysis Group, Bioinformatics Institute, A*STAR, Singapore
sSchool of Computing, National University of Singapore, Singapore
tBeijing Shanggang Medical Technology Co., Ltd., China
uDepartment of Computer Science and Engineering, Shanghai Jiao Tong University, China
vMoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China
wJ. Crayton Pruitt Family Department of Biomedical Engineering, University of Florida, USA
xIndian Institute of Technology Kharagpur, India
yINESC TEC - Institute for Systems and Computer Engineering, Technology and Science, Porto, Portugal
zFEUP - Faculty of Engineering of the University of Porto, Porto, Portugal
aaINSEM, UMR 1101, Brest, France
abDepartment of Electrical and Electronic Engineering, Universiti Teknologi PETRONAS, Malaysia
acInViA/IFTIM, Université de Bourgogne, Dijon, France

*Corresponding author
Email address: porwal.prasanna@sggs.ac.in (Prasanna Porwal)
These authors co-organized the challenge. All others contributed results of their algorithm(s) presented in the paper

Preprint submitted to Medical Image Analysis June 10, 2019
Abstract

Diabetic Retinopathy (DR) is the most common cause of avoidable vision loss, predominantly affecting the working age population across the globe. Screening for DR, coupled with timely consultation and treatment, is a globally trusted policy to avoid vision loss. However, the implementation of DR screening programs is challenging due to the scarcity of medical professionals able to screen a growing global diabetic population at risk for DR. Computer-aided disease diagnosis in retinal image analysis could provide a sustainable approach for such large-scale screening effort. The recent scientific advances in computing capacity and machine learning approaches provide an avenue for biomedical scientists to reach this goal. Aiming to advance the state-of-the-art in automatic DR diagnosis, the Grand Challenge on “Diabetic Retinopathy Segmentation and Grading” was organized in conjunction with the IEEE International Symposium on Biomedical Imaging (ISBI - 2018). In this paper, we report the set-up and results of this challenge that is primarily based on Indian Diabetic Retinopathy Image Dataset (IDRiD). There were three principal sub-challenges: lesion segmentation, disease severity grading, and localization of retinal landmarks and segmentation. These multiple tasks in this challenge allow to test the generalizability of the algorithms, and this is what makes it different from the existing ones. It received a positive response from a scientific community with 148 submissions from 495 registrations effectively entered in this challenge. This paper outlines the challenge, its organization, the dataset used, evaluation methods and results of top performing participating solutions. We observe that the top performing approaches utilize a blend of clinical information, data augmentation, and the ensemble of models. These findings have the potential to enable new developments in retinal image analysis and image-based DR screening in particular.

Keywords: Diabetic Retinopathy; Retinal image analysis; Deep learning; Challenge
1. Introduction

Diabetic Retinopathy (DR) and Diabetic Macular Edema (DME) are the most common sight-threatening medical conditions caused due to retinal microvascular changes triggered by diabetes [Reichel and Salz 2015], predominantly affecting the working-age population in the world [Atlas 2017]. DR leads to gradual changes in the vasculature structure (including vascular tortuosity, branching angles and calibers) and resulting abnormalities (microaneurysms, haemorrhages and exudates), whereas, DME is characterized by the retention of fluid or swelling of macula that may occur at any stage of DR [Bandello et al. 2010; Ciulla et al. 2003]. According to the International Diabetes Federation [Atlas 2017] estimates, presently, the global number individuals affected with diabetes is 425 million, and it may rise to 693 million by 2045. Amongst them, 1 out of 3 individuals are estimated to have some form of DR and 1 in 10 is prone to vision-threatening DR [ICO 2017; Bourne et al. 2013]. DR is diagnosed by visually inspecting retinal fundus images for the presence of one or more retinal lesions like microaneurysms (MAs), hemorrhages (HEs), soft exudates (SEs) and hard exudates (EXs) [Wong et al. 2016] as shown in Fig. 1.

![Fig. 1. Illustration of retinal image (in center) by highlighting normal structures (blood vessels, optic disc and fovea center) and abnormalities associated with DR: Enlarged regions (in left) MAs, and HEs and (in right) SEs, and EXs.](image)

Early diagnosis and treatment of DR can prevent vision loss. Hence, diabetic patients are typically referred for retinal screening once or twice a year [Ferris 1993].
on the number of ophthalmologists and necessary health care infrastructure (Jones and Edwards, 2010; Lin et al., 2016). In the Indian subcontinent, ophthalmologist to population ratio is 1:107,000, however, in urban regions this ratio is 1:9000 whereas in rural parts there is only one ophthalmologist for 608,000 inhabitants (Raman et al., 2016). By 2045, India alone is projected to have approximately 151 million people with diabetes and one-third of them are expected to have DR (Atlas, 2017). Programs to screen such a large population for DR confront the issues related to the implementation, management, availability of human graders, and long-term financial sustainability. Hence, computer aided diagnosis tools are required for screening such a large population that require continuous follow-up for DR and to effectively facilitate in reducing the burden on the ophthalmologists (Jelinek and Cree, 2009; Walter et al., 2002). Such a tool would help clinicians in the identification, interpretation, and measurements of retinal abnormalities, and ultimately in the screening and monitoring of the disease. The recent scientific advances in computing capacity and machine learning approaches provide an avenue to the biomedical scientists to meet the desideratum of clinical practice (Shortliffe and Blois, 2006; Patton et al., 2006). To meet this need raw images along with precise pixel or image level expert annotations (also known as ground truths) play an important role to facilitate the research community for the development, validation, and comparison of DR lesion segmentation techniques (Trucco et al., 2013). Precise pixel-level annotations of lesions associated with DR such as MAs, HEs, SEs and EXs are invaluable resource for evaluating accuracy of individual lesion segmentation techniques. These precisely segmented lesions help in determining the disease severity and further act as a road-map that can assist to tap the progression of disease during follow-up procedures. Similarly, on the other hand, image-level expert labels for disease severity of DR, and DME are helpful in the development and evaluation of image analysis and retrieval algorithms. This necessity has led several research groups to develop and share retinal image datasets, namely Messidor (Decenci`ere et al., 2014), Kaggle (Cuadros and Bresnick, 2009), ROC (Niemeijer et al., 2010), E-Ophth (Decenci`ere et al., 2013), DiaretDB (Kauppi et al., 2012), DRIVE (Staal et al., 2004), STARE (Hoover, 1975), ARIA (Farnell et al., 2008) and HEI-MED (Giancardo et al., 2016).
Further, two challenges were organized in the context of DR, namely Retinopathy Online Challenge (ROC)\(^2\) and Kaggle DR detection challenge\(^3\). ROC was organized with the goal of detecting MAs. Whereas, the Kaggle challenge aimed to get solution for determining the severity level of DR. These challenges enabled advances in the field by promoting the participation of scientific research community from all over the globe on a competitive at the same time constructive setting for scientific advancement. Previous efforts have made good progress using image classification, pattern recognition, and machine learning. The progress through last two decades has been systematically reviewed by several research groups (Patton et al., 2006; Winder et al., 2009; Abràmoff et al., 2010; Mookiah et al., 2013a; Jordan et al., 2017; Nørgaard and Grauslund, 2018).

Although lots of efforts have been made in the field towards automating the DR screening process, lesion detection is still a challenging task due to the following aspects: (a) Complex structures of the lesions (shape, size, intensity), (b) detection of lesions in tessellated images and in presence of noise (bright border reflections, impulsive noise, optical reflections), (c) high inter-class similarity (i.e. between MA-HE and EX-SE), (d) appearance of not so uncommon non-lesion structures (nerve fiber reflections, vessel reflections, drusen) and (e) difference in images obtained by different imaging devices makes it difficult to build a flexible and robust model for lesion segmentation. To the best of our knowledge, prior to the challenge, there were no reports on the development of a single framework to segment all lesions (MA, HE, SE, and EX) simultaneously. Also, there was a lack of common platform to test the robustness of approaches that determine the normal and abnormal retinal structures on the same set of images. Furthermore, there was limited availability of the pixel level annotations and the simultaneous gradings for DR and DME (see Tables in Appendix A).

In order to address these issues, we introduced a new dataset called Indian Diabetic Retinopathy Image Dataset (IDRiD) (Porwal et al., 2018a). Further, it was used as a base dataset for the organization of grand challenge on “Diabetic Retinopathy: Seg-

\(^2\)http://webeye.ophth.uiowa.edu/ROC/
\(^3\)https://www.kaggle.com/c/diabetic-retinopathy-detection
mentation and Grading” in conjunction with ISBI - 2018. The IDRiD dataset provides expert markups of typical DR lesions and normal retinal structures. It also provides disease severity level of DR, and DME for each image in the database. This challenge brought together the computer vision and biomedical researchers with an ultimate aim to further stimulate and promote research, as well as to provide a unique platform for the development of a practical software tool that will support efficient and accurate measurement and analysis of retinal images that could be useful in DR management. Initially, a training dataset along with the ground truth was provided to participants for the development of their algorithms. Later, the results were judged on the performance of these algorithms on test dataset. Success was measured by how closely the algorithmic outcome matched the ground truth. There were three principal sub-challenges: lesion segmentation, disease severity grading, and localization and segmentation of retinal landmarks. These multiple tasks in IDRiD challenge allow to test the generalizability of the algorithms, and this is what makes it different from the existing ones. Further, this challenge seeks an automated solution to predict the severity of DR and DME simultaneously. It was projected as an individual task to increase the difficulty level of this challenge as compared to the Kaggle DR challenge i.e. for a given image, the predicted severity for both DR and DME should be correct to count for scoring the task.

The rest of the paper is structured as follows: Section 2 gives a short review of previous work done in the development of automated DR screening, section 3 provides details of reference dataset, section 4 describes the organization of the competition through various phases and section 5 details the top performing competing solutions. Section 6 presents performance evaluation measures used in this challenge. Then, section 7 presents the results, analysis and corresponding ranking of participating teams for all sub-challenges. Section 8 provides a brief discussion on the results, limitations, and lessons learned from this challenge and at last the conclusion. Along with this the paper, Appendix A is included that provides a comparison of different state-of-the-art publicly available databases with the IDRiD dataset.
2. Review of Retinal Image Analysis for the detection of DR

Automatic image processing has proven to be a promising choice for the analysis of retinal fundus images and its application to future eye care. The introduction of automated techniques in DR screening programs and the interesting outcomes achieved by the rapidly growing deep learning technology are examples of success stories and potential future achievements. Particularly, after researcher's (Krizhevsky et al., 2012) deep learning based model showed significant improvements over the state of the art in the ImageNet challenge, there was a surge of deep learning based models in medical image analysis. Hence, we decided to present the most recent relevant works with a classification based on whether or not they used deep learning in the context of DR.

2.1. Non-deep learning methods

The general framework for retinal image analysis through traditional handcrafted features based approaches involve several stages, typically: a preprocessing stage for contrast enhancement or non-uniformity equalization, image segmentation, feature extraction, and classification. The feature extraction strategy varies according to the objective involved i.e. retinal lesion detection, disease screening or landmark localization. In 2006, one research group (Patton et al., 2006) outlined the principles upon which retinal image analysis is based and discussed the initial techniques used to detect the retinal landmarks and lesions associated with DR. Later, one another group (Winder et al., 2009) reported an analysis of the work in the automated analysis of DR during 1998–2008. They categorized the literature into a series of operations or steps as preprocessing, vasculature segmentation, localization, and segmentation of the optic disk (OD), localization of the macula and fovea, detection and segmentation of lesions. Some of the review articles (Abramoff et al., 2010; Jordan et al., 2017) provide a brief introduction to quantitative methods for the analysis of fundus images with a focus on identification of retinal lesions and automated techniques for large scale screening for retinal diseases. Majority of attempts in the literature are towards exclusive detection and/or segmentation of one type of lesions (either MAs, HEs, EXs or SEs) from an image. Some of the common approaches involved for lesion segmentation are mathematical morphology (Joshi and Karule, 2019; Hatanaka et al., 2008)
Zhang et al., 2014), region growing (Fleming et al., 2006; Li and Chutatape, 2004), and supervised (Wu et al., 2017; Zhou et al., 2017; Garcia et al., 2009; Tang et al., 2013). Apart from these approaches, in case of MAs, most initial studies shown the effectiveness of template matching (Quellec et al., 2008), entropy thresholding (Das et al., 2015), radon space (Giancardo et al., 2011), sparse representation (Zhang et al., 2012; Javidi et al., 2017), hessian based region descriptors (Adal et al., 2014), dictionary learning (Rocha et al., 2012). On the other hand, for exclusive segmentation of HEs, super-pixel based features (Tang et al., 2013; Romero-Oraá et al., 2019) were found to be effective. These red lesions (both MAs and HEs) are also frequently detected together using dynamic shape features (Seoud et al., 2016), filter response and multiple kernel learning (Srivastava et al., 2017) and hybrid feature extraction approach (Niemeyer et al., 2005). Similarly, for EXs researchers relied on approaches like clustering (Osareh et al., 2009), model-based (Sánchez et al., 2009; Harangi and Hajdu, 2014), ant colony optimization (ACO) (Pereira et al., 2015) and contextual information (Sánchez et al., 2012). Whereas, for SEs researchers utilized Scale Invariant Feature Transform (SIFT) (Naqvi et al., 2018), adaptive thresholding and ACO (Sreng et al., 2019). Further, several approaches were devised for multiple lesion detection such as multiscale amplitude-modulation-frequency-modulation (Agurto et al., 2010), machine learning (Roychowdhury et al., 2014), a combination of Hessian multiscale analysis, variational segmentation and texture features (Figueiredo et al., 2015). These techniques are shown to usually involve interdependence on the detection of anatomical structures (i.e. OD and fovea) with the lesion detection, and that in turn determines the automated DR screening outcome.

Localization and segmentation of OD and fovea facilitate the detection of retinal lesions as well as in the assessment (based on the geometric location of these lesions) of the severity and monitoring the progression of DR and DME. Hence, several approaches have been proposed for localization of OD, most of them utilized the OD properties like intensity, shape, color, texture, etc. and many others showed the effectiveness of mathematical morphology (Morales et al., 2013; Marin et al., 2015), template matching (Giachetti et al., 2014), deformable models (Yu et al., 2012; Wu et al., 2016) and intensity profile analysis (Kamble et al., 2017; Uribe-Valencia and
Martinez-Carballido (2019). Further, the approaches utilized for OD segmentation are based on level set (Yu et al., 2012), thresholding (Marin et al., 2015), active contour (Mary et al., 2015) and shape modeling (Cheng et al., 2015), clustering (Thakur and Juneja, 2017), and hybrid (Bai et al., 2014) approaches. Similarly, the fovea is detected mostly using the geometric relationship with OD and vessels through morphological (Welfer et al., 2011), thresholding (Gegundez-Arias et al., 2013), template (Kao et al., 2014) and intensity profile analysis (Kamble et al., 2017) techniques. Poor performance on detection of the normal anatomical structures could adversely affect lesion detection and screening accuracy. For instance, consider the mathematical morphology based techniques presented in 2002 (Walter et al., 2002), 2008 (Sopharak et al., 2008) and 2014 (Zhang et al., 2014). These works demonstrate how the morphological processing-based approaches evolved by including multiple steps for the final objective of exudate detection. In the initial efforts, Walter et al. devised a technique for OD and EXs segmentation, later removed the OD to obtain the exudate candidates. Similarly, Sopharak et al. achieved the same objective with the detection, and removal of OD and vessels. Recently, the approach presented by Zhang et al. achieved much better result, but it involved (a) spatial calibration, (b) detection of dark and bright anatomical structures such as vessels and OD respectively, also (c) bright border regions detection before actual extraction of candidates. Also, there are other techniques based on textual (Morales et al., 2017; Porwal et al., 2018c) and mid-level (Pires et al., 2017) features of retinal images that forgo the lesion segmentation step for DR screening. However, most of these techniques depend on the intermediate steps mentioned above. In the approach based on machine learning (Roychowdhury et al., 2014) detected bright and dark lesions as a first step and later performed the hierarchical lesion classification to generate a severity grade for DR. Similarly, Antal and Hajdu (2014) proposed a strategy involving image-level quality assessment, pre-screening followed by lesion and anatomical features extraction to finally decide about the presence of DR using ensemble of classifiers. Further, for identification of different stages of DR features from morphological region properties (Yun et al., 2008), texture parameters (Acharya et al., 2012; Mookiah et al., 2013b), non-linear features of the higher-order spectra (Acharya et al., 2008), hybrid (Dhara et al., 2015) and information fusion (Niemeyer et al., 2009).
approaches were found useful. As the DME is graded based on the location of the EXs from macula, many researchers (Giancardo et al., 2012; Medhi and Dandapat, 2014; Perdomo et al., 2016; Marin et al., 2018) proposed EXs based features to determine the severity of the DME. While several others (Deepak and Sivaswamy, 2012; Mookiah et al., 2015; Acharya et al., 2017) have proposed various feature extraction techniques to grade DME stages without segmenting EXs. Mainly for the approaches in this section, the features are based on the color, brightness, size, shape, edge strength, texture, and contextual information of pixel clusters in spatial and/or transform domain. Whereas the classification is achieved through the classifiers such as K Nearest Neighbors (KNN), Naive Bayes, Support Vector Machine (SVM), Artificial Neural Network (ANN), Decision Trees, etc.

These lesion detection or screening techniques are shown to usually involve interdependence with the other landmark detection. However, there is a lack of single platform to test their performance for each objective. For such handcrafted features based approaches this challenge provides a unique platform to compare and contrast the algorithm’s performance for the detection of anatomical structures, lesions as well as screening of DR and DME.

2.2. Deep learning methods

Deep Learning is a general term to define multi-layered neural networks able to concurrently learn a low-level data representation and higher-level parameters directly from the data. This representation learning capability drastically reduces the need for engineering ad-hoc features, however, the full end-to-end training of deep learning-based approaches typically require a significant number of samples. Its rapid development in recent times is mostly due to a massive influx of data, advances in computing power and developments in learning algorithms that enabled the construction of multi-layer (more than two) networks (Hinton, 2018; Voulodimos et al., 2018). This progress has induced interests in the creation of analytical, data-driven models based on machine learning in health informatics (Ching et al., 2018; Ravi et al., 2017). Hence, it is emerging as an effective tool for machine learning, promising to reshape the future of automated medical image analysis (Greenspan et al., 2016; Litjens et al., 2017; Suzuki, 2018).
Among various methodological variants of deep learning, Convolutional Neural Networks (CNNs or ConvNets) are the most popular within the field of medical image analysis (Hoo-Chang et al., 2016). Several configurations and variants of CNN’s are available in the literature, some of the most popular are AlexNet (Krizhevsky et al., 2012), VGG (Simonyan and Zisserman, 2014), GoogLeNet (Szegedy et al., 2015) and ResNet (He et al., 2016).

Deep learning has also been widely utilized in the retinal image analysis because of its unique characteristic of preserving local image relations. Majority of the approaches in the literature employ deep learning to retinal images by utilizing “off-the-shelf CNN” features as complementary information channels to other handcrafted features or local saliency maps for detection of abnormalities associated with DR (Chudzik et al., 2018; Orlando et al., 2018; Dai et al., 2018), segmentation of OD (Zilly et al., 2017; Fu et al., 2018), and the detection of DR (Rangrej and Sivaswamy, 2017). The authors (Fu et al., 2016) employ fully connected conditional random fields along with CNN to integrate the discriminative vessel probability map and long-range interactions between pixels to obtain final binary vasculature. Whereas some approaches initialized the parameters with those of pre-trained models (on non-medical images), then “fine-tuned” (Tajbakhsh et al., 2016) the network parameters for DR screening (Gulshan et al., 2016; Carson Lam et al., 2018). In another approach researchers used two-dimensional (2D) image patches as an input instead the full-sized images for lesion detection (Tan et al., 2017b; van Grinsven et al., 2016; Lam et al., 2018; Chudzik et al., 2018; Khojasteh et al., 2018), and OD and fovea detection (Tan et al., 2017a). In Garcia et al. (2017) trained the “CNN from scratch” and compared it with the fine-tuning results based on the other two existing architectures. Recently, Shah et al. (2018) demonstrated that the ensemble training of auto-encoders stimulates diversity in learning dictionary of visual kernels for detection of abnormalities. Whereas Giancardo et al. (2017) proposed a novel way to compute the vasculature embedding that leverages the internal representation of a new encoder-enhanced CNN, demonstrating improvement in the DR classification and retrieval task.

There is a significant development in the automated identification of DR using CNN
models in recent time. A customized CNN (Gargeya and Leng, 2017) proposed for DR screening and trained using 75,137 obtained from EyePACS system (Cuadros and Bresnick, 2009), where an additional classifier was further employed on the CNN-derived features to determine if the image is with or without retinopathy. Similarly, Google Inc. (Gulshan et al., 2016) developed a network optimized (fine tuning) for image classification, in which a CNN is trained by utilizing a retrospective development database consisting of 128,175 images with the labels. There are some hybrid algorithms, in which multiple, semi-dependent CNN’s are trained based on the appearance of retinal lesions (Abramoff et al., 2016; Quellec et al., 2016). A step further, the researchers (Quellec et al., 2017) demonstrated an ability of lesion segmentation based on the CNN trained for image level classification. However, Lynch et al. (2017) demonstrated that the hybrid algorithms based on multiple semi-dependent CNNs might offer a more robust option for DR referral screening, stressing the importance of lesion segmentation. For further details, readers are recommended to follow recent reviews for detection of exudates (Fraz et al., 2018), red lesions (Biyani and Patre, 2018) and a systematic review with a focus on the computer-aided diagnosis of DR (Mookiah et al., 2013a; Nørgaard and Grauslund, 2018).

This current progress in artificial intelligence provides an opportunity to the researchers for enhancing the performance of the DR referral system to more robust diagnosis system that can provide the quantitative information for multiple diseases matching the international standards of clinical relevance. Thus, this challenging design offers an avenue to gauge precise DR severity status and opportunity to deliver accurate measures for lesions, that could even help in the follow-up studies to observe changes in the retinal atlas.

3. Indian Diabetic Retinopathy Image Dataset

3.1. Image Acquisition

The IDRiD dataset (Porwal et al., 2018a) was created from real clinical exams acquired at an Eye Clinic located in Nanded, (M.S.), India. The fundus photographs of people affected by diabetes were captured with focus on macula using Kowa VX-10α.
fundus camera. Prior to capturing of images, pupils of all subjects were dilated with one drop of tropicamide at 0.5% concentration. The captured images have $50^\circ$ field of view and resolution of $4288 \times 2848$ pixels stored in *jpg* format. The final dataset is composed of 516 images divided into five DR ($0 - 4$) and three DME ($0 - 2$) classes with well-defined characteristics according to international standards of clinical relevance. It provides expert markups of typical diabetic retinopathy lesions and normal retinal structures. It also provides disease severity level of DR, and DME for each image in the database. Three types of ground-truths are available in the dataset:

1. **Pixel Level Annotations.** This type of annotations are useful in the techniques to locate individual lesions within an image and to segment out regions of interest from the background. Eighty-one color fundus photographs with signs of DR are annotated at pixel level for developing ground truth of MAs, SEs, EXs and HEs. The binary masks (as shown in Fig. 2) for each type of lesion are provided in tif file format. Additionally, OD was also annotated at pixel level and binary masks for all 81 images are provided in the same format. These annotations play a vital role in the research for the computational analysis of segmenting lesions within the image.

![Fig. 2. Retinal photograph and different annotations: (a) sample fundus image from the IDRiD dataset; sample ground truths of (b-f) MAs, HEs, SEs, EXs and OD respectively.](image)

---

13
2. Image Level Grading. It consists of information meant to describe overall risk factors associated with an entire image. Two medical experts graded the full set of 516 images with a variety of pathological conditions of DR and DME. Grading for all images is available in CSV file. The diabetic retinal images were classified into separate groups according to the International Clinical Diabetic Retinopathy Scale (Wu et al., 2013) as shown in Table 1. The DME severity was decided based on occurrences of EXs near to macula center region (Decenci`ere et al., 2014) as shown in Table 2.

### Table 1. DR Severity Grading.

<table>
<thead>
<tr>
<th>DR Grade</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>0: No apparent retinopathy</td>
<td>No visible sign of abnormalities</td>
</tr>
<tr>
<td>1: Mild – NPDR</td>
<td>Presence of MAs only</td>
</tr>
<tr>
<td>2: Moderate – NPDR</td>
<td>More than just MAs but less than severe NPDR</td>
</tr>
<tr>
<td>3: Severe – NPDR</td>
<td>Any of the following: &gt;20 intraretinal HEs</td>
</tr>
<tr>
<td></td>
<td>Venous beading</td>
</tr>
<tr>
<td></td>
<td>Intraretinal microvascular abnormalities</td>
</tr>
<tr>
<td></td>
<td>no signs of PDR</td>
</tr>
<tr>
<td>4: PDR</td>
<td>Either or both of the following: Neovascularization</td>
</tr>
<tr>
<td></td>
<td>Vitreous/pre-retinal HE</td>
</tr>
</tbody>
</table>

### Table 2. Risk of DME.

<table>
<thead>
<tr>
<th>DME Grade</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>No Apparent EX(s)</td>
</tr>
<tr>
<td>1</td>
<td>Presence of EX(s) outside the radius of one disc diameter from the macula center</td>
</tr>
<tr>
<td>2</td>
<td>Presence of EX(s) within the radius of one disc diameter from the macula center</td>
</tr>
</tbody>
</table>

3. Optic Disc and Fovea center co-ordinates. The OD and fovea center locations are marked for all 516 images and the markup is available as separate CSV file.
The IDRiD dataset is available from the IEEE Dataport Repository\footnote{https://ieee-dataport.org/open-access/indian-diabetic-retinopathy-image-dataset-idrid} under a Creative Commons Attribution 4.0 License. The more detailed information about the data is available in the data descriptor \cite{Porwal2018b}. Tables A.1 and A.2 highlight a comparative strength of the presented dataset with respect to the existing datasets. IDRiD is the only dataset that provides all three types of annotations mentioned above. Streamlining the collection of annotations would allow it to be utilized in research and would lead to better generalizable models for image analysis to be developed, enabling further progress in the automated DR diagnosis.

## 4. Challenge Organization

The “Diabetic Retinopathy: Segmentation and Grading” challenge was composed into various stages, giving a well-organized work process to potentiate the success of the contest. Fig. 3 depicts the work-flow of the overall challenge organization. The challenge was officially announced at the ISBI - 2018 website\footnote{https://biomedicalimaging.org/2018/challenges/} on 15\textsuperscript{th} October 2017. The challenge was subdivided into three sub-challenges as follows:

1. **Lesion Segmentation**: Segmentation of retinal lesions associated with DR as MAs, HEs, EXs and SEs.

2. **Disease Grading**: Classification of fundus images according to the severity level of DR and DME.

3. **OD detection and Segmentation, and Fovea Detection**: Automatic localization of OD and fovea center coordinates, and segmentation of OD.

The challenge involved 4 stages, as detailed below:

### Stage 1: Data Preparation and Distribution

The IDRiD dataset was adopted for this challenge, where experts verified that all images are of adequate quality, clinically relevant, that no image is duplicated and that a reasonable mixture of disease stratification representative of DR and DME is present. The dataset along with the ground truths were separated into training set and test set. For the images with pixel level annotations,
the data was separated as 2/3 for training (Set-A) and 1/3 for testing (Set-B) (See Table 3). Similarly, data for the OD segmentation (part of sub-challenge – 3) was divided in the same ratio into Set-A (54 images) and Set-B (27 images). The percentage of images that should be in each subset for lesion and OD segmentation tasks (sub-challenge – 1 and part of sub-challenge – 3) were chosen based on the research outcome [Dobbin and Simon, 2011] which demonstrated that splitting data into 2/3 (training): 1/3 (testing) is an optimal choice for the sample sizes from 50 to 200. For the other sub-challenges (disease grading, and OD and fovea center locations), data was separated in 80 (training set: Set-A): 20 (testing set: Set-B) ratio. The percentage of data split in this case is

<table>
<thead>
<tr>
<th>Lesion Type</th>
<th>Set - A Images</th>
<th>Set - B Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>MA</td>
<td>54</td>
<td>27</td>
</tr>
<tr>
<td>HE</td>
<td>53</td>
<td>27</td>
</tr>
<tr>
<td>SE</td>
<td>26</td>
<td>14</td>
</tr>
<tr>
<td>EX</td>
<td>54</td>
<td>27</td>
</tr>
</tbody>
</table>

Table 3. Stratification of retinal images annotated at pixel level for different types of retinal lesions.
done to provide an adequate amount of data divided into different severity levels. Note that the dataset was stratified according the DR and DME grades before splitting. A breakdown of the details of the dataset is shown in Table

<table>
<thead>
<tr>
<th>DR Grade</th>
<th>Set-A</th>
<th>Set-B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>134</td>
<td>34</td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>136</td>
<td>32</td>
</tr>
<tr>
<td>3</td>
<td>74</td>
<td>19</td>
</tr>
<tr>
<td>4</td>
<td>49</td>
<td>13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DME Grade</th>
<th>Set-A</th>
<th>Set-B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>177</td>
<td>45</td>
</tr>
<tr>
<td>1</td>
<td>41</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>195</td>
<td>48</td>
</tr>
</tbody>
</table>

The challenge was hosted on *Grand Challenges in Biomedical Imaging Platform*[^6], one of the popular platform for biomedical imaging-related competitions. A challenge website was set up and launched on 25\textsuperscript{th} October 2017 to disseminate challenge related information. It was also used for registration, data distribution, submission of results and paper, and communication between the organizers and participants.

**Stage 2: Registration and release of the training data.** The registration of challenge for consideration to ISBI on-site contest was open from the launch of grand-challenge website (i.e. 25\textsuperscript{th} October 2017) till deadline for the submission of results (i.e. 11\textsuperscript{th} March 2018). Interested research teams could register through challenge website for one or all sub-challenges. The first part of data, Set-A (images and ground truths) was made available to participants of the challenge on 20\textsuperscript{th} January 2018. Participants could download the dataset and start development or modification of their methods. Further, they were also allowed to use other datasets for the development of their methods, with the condition that the external datasets be publicly available.

**Stage 3: Release of test data.** The Set-B (only images) for sub-challenge – 1 was released on 20\textsuperscript{th} February, 2018. For other two sub-challenges, the Set-B was released on 4\textsuperscript{th} April which was part of “on-site” challenge. The organizers refrained from an

[^6]: https://grand-challenge.org/
on-site evaluation of sub-challenge – 1 considering the timing constrains in evaluation of the results for individual image segmentation results.

Submissions were sought for either of the following 8 different tasks corresponding to the three sub-challenges (1 – Lesion Segmentation, 2 – Disease Grading, 3 – OD and Fovea Detection) as follows:

1. Sub-challenge – 1: Lesion Segmentation
   - Task - 1: MA Segmentation
   - Task - 2: HE Segmentation
   - Task - 3: SE Segmentation
   - Task - 4: EX Segmentation
2. Sub-challenge – 2: Disease Grading
   - Task - 5: DR and DME Grading
3. Sub-challenge – 3: Optic Disc and Fovea Detection
   - Task - 6: OD Center Localization
   - Task - 7: Fovea Center Localization
   - Task - 8: OD Segmentation

Challenge site was made open for submission from 12th February and participants could submit their results and paper describing their approach till March 11, 2018 to the organizers. Participants could submit up to three methods to be evaluated per team for each task, provided that there was a significant difference between the techniques, beyond a simple change or alteration of parameters. For Tasks 1 to 4 (i.e. sub-challenge – 1) and task-8, the teams were asked to submit output probability maps as grayscale images and for all other tasks it was accepted in CSV format. The submitted results were evaluated by the challenge organizers and their performance was displayed on leaderboard of the challenge website. For sub-challenge – 1, the teams were assessed based on the performance of results submitted on the test set, whereas, for other two sub-challenges assessment was based on the results on the training set obtained through leave one out cross-validation approach. In this phase, it received very good response from the research community with 148 submissions by 37 different teams, out of which 16 teams were shortlisted for participation to the on-site challenge. Amongst invited,
13 teams confirmed their participation in the on-site challenge, whereas, two teams declined to participate due to other commitments and one team was not able arrange financial support in the limited time.

Stage 4: ISBI Challenge Event. The main challenge event was held in conjunction with ISBI - 2018 on April 4th, 2018. The Set-B (only images) for sub-challenge – 2 and 3 was made available to the participants via challenge website (on-line mode) as well as portable devices at the challenge site (off-line mode). Participants were asked to produce results for respective challenge task within one hour. The participating teams could bring their own system or run the test through the remote system. Also, there was no restriction on the number of machines that could be used to produce the results. However, considering the timing constraints for processing, some teams which had previously entered with more than one solution decided to use only their best performing solution.

Further, the top three teams from sub-challenge – 1 were given opportunity to present their work. During that time, some of the organizing team members compiled the results for sub-challenge – 2 and 3. The teams were given 7 minutes for presentation of their approach and 3 minutes were reserved for question-answers. The first presentation session lasted for about 30 minutes and at the end of presentations of sub-challenge – 1 the result for sub-challenge – 2 and 3 were declared. Similarly, the top three performing teams from these sub-challenges gave short presentations on their work. After the end of the on-site challenge event, on 6th April, the summary of challenge and analysis of results were presented, which included a final ranking of the competing solutions. This information is additionally accessible on the challenge website. It is important to note that many teams had participated in multiple sub-challenges as listed in the Table 5 and remainder of this paper deals only with the methods that were selected for the challenge.

5. Competing Solutions

Majority of participating teams proposed a CNN based approach for solving tasks in this challenge. This section details the basic terminologies and abbreviations related
Table 5. List of all participating teams shortlisted and which participated in the ‘on-site’ challenge. All teams are color coded for easier reference in all further listings. The DL denotes whether the submitted algorithm is based on deep learning. Where, sub-challenge – 1 (SC1) corresponds to lesion segmentation such as microaneurysms (MA), haemorrhages (HE), soft exudates (SE) and hard exudates (EX). Whereas, sub-challenge – 2 (SC2) denotes disease severity grading corresponding to DR and DME. Similarly, sub-challenge – 3 (SC-3) deals with the optic disc detection (ODD), fovea detection (FD) and optic disc segmentation (ODS). Harangi et al. participated with two methods HarangiM1 and HarangiM2, for simplicity it is jointly represented as HarangiM1-M2 with a single color code. Similarly, Li et al. participated with two methods LzyUNCC (renamed in text as LzyUNCC-I) and LzyUNCC_Fusion (renamed in text as LzyUNCC-II) that are jointly represented as LzyUNCC with same color code. However, these different methods are mentioned separately in the text wherever it was necessary. *Team could not participate in ‘on-site’ challenge but later communicated the results to the organizers.

<table>
<thead>
<tr>
<th>Team Name</th>
<th>Authors</th>
<th>DL</th>
<th>SC1 MA</th>
<th>SC1 HE</th>
<th>SC1 SE</th>
<th>SC1 EX</th>
<th>SC2 ODD</th>
<th>SC2 FD</th>
<th>SC2 ODS</th>
</tr>
</thead>
<tbody>
<tr>
<td>VRT</td>
<td>Jaemin Son et al.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>iFLYTEK-MIG</td>
<td>Fengyan Wang et al.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>PATech</td>
<td>Liu Lihong et al.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>SOONER</td>
<td>Yunzhi Wang et al.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>SAIHST</td>
<td>Yoon Ho Choi et al.</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>LzyUNCC</td>
<td>Zhongyu Li et al.</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>SDNU</td>
<td>Xiaodan Sui et al.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Mammoth</td>
<td>Junyan Wu et al.</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>HarangiM1-M2</td>
<td>Balazs Harangi et al.</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>AVSASVA</td>
<td>Varghese Alex et al.</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>DeepDR</td>
<td>Ling Dai et al.</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>ZJU-Bll-SGEX</td>
<td>Xingzheng Lyu et al.</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>IITkgpKLIV</td>
<td>Oindrila Saha et al.</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>*CBER</td>
<td>Ana Mendonça et al.</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

For the input image, CNN transforms the raw image pixels on one end to generate a single differentiable score function at the other. It exploits three mechanisms — sparse connections (a.k.a. local receptive field), weight sharing and invariant (or equivariant) representation — that makes it computationally efficient (Shen et al., 2017). The CNN architecture typically consists of an input layer followed by sequence of convolutional
(CONV), subsampling (POOL), fully-connected (FC) layers and finally a Softmax or regression layer, to generate the desired output. Functions of all layers are detailed as follows:

The CONV layer comprises of a set of independent filters (or kernels) that are utilized to perform 2D convolution with the input layer \( I \) to produce the feature (or activation) maps \( A \) that give the responses of kernels at every spatial position. Mathematically, for the input patch \( I^\ell_{x,y} \) centered at location \((x, y)\) of the \( \ell \)th layer, the feature value in the \( i \)th feature map, \( A^\ell_{x,y,i} \), is obtained as:

\[
A^\ell_{x,y,i} = f((w^\ell_i)^T I^\ell_{x,y} + b^\ell_i) = f(C^\ell_{x,y,i}) \quad (1)
\]

Where the parameters \( w^\ell_i \) and \( b^\ell_i \) are weight vector and bias term of the \( i \)th filter of the \( \ell \)th layer, and \( f(\cdot) \) is a nonlinear activation function such as sigmoid, rectified linear unit (ReLU) or hyperbolic tangent (tanh). It is important to note that the kernel \( w^\ell_i \) that generates the feature map \( C^\ell_{x,y,i} \) is shared, reducing the model complexity and making the network easier to train.

The POOL layer aims to achieve translation-invariance by reducing the resolution of the feature maps. Each unit in a feature map of the POOL layer is derived using a subset of units within sparse connections from the corresponding convolutional feature map. The most common pooling operations are average pooling and max pooling. It performs downsampling operation and is usually placed between two CONV layers to achieve a hierarchical set of image features. The kernels in the initial CONV layers detect low-level features such as edges and curves, while the kernels in the higher layers are learned to encode more abstract features. The sequence of several CONV and POOL layers gradually extract higher-level feature representation.

FC layer aims to perform higher-level reasoning by computing the class scores. Each neuron in this layer is connected to all neurons in the previous layer to generate global semantic information.

The last layer of CNN’s is an output layer \( O \), here the Softmax operator is commonly used for the classification tasks. The optimum parameters \( \theta \) (notation for both \( w \) and \( b \)) for a particular task can be determined by minimizing the
loss function \( (L) \) defined for the task. Mathematically, for \( N \) input-output relations 
\[
\{(I^n, O^n); n \in [1, \cdots, N]\}
\] 
and corresponding labels \( G^n \) the loss can be derived as:

\[
L = \frac{1}{N} \sum_{n=1}^{N} \ln(\theta; G^n, O^n) 
\]

(2)

Where \( N \) denotes the number of training images, \( I^n, O^n \) and \( G^n \) correspond to 
the \( n^{th} \) training image. Here, a critical challenge in training CNN’s arises from the 
limited number of training samples as compared to the number of learnable parameters 
that need to be optimized for the task at hand. Recent studies have developed some 
key techniques to better train and optimize the deep models such as data augmentation, 
weight initialization, Stochastic Gradient Descent (SGD), batch normalization, 
shortcut connections and regularization. For more understanding related to advances 
in CNN’s, reader is recommended to refer (Gu et al., 2018).

The growing use of CNN’s as the backbone of many visual tasks, ready for different 
purposes (such as segmentation, classification or localization) and available data, has 
made architecture search a primary channel in solving the problem.

In this challenge, mainly for disease severity grading problem, participants either 
directly utilized existing variants of CNN’s or ensembled them to demarcate the input 
to one of the class mentioned above. Several configurations and variants of 
CNN’s are available in literature, some of the most popular are AlexNet (Krizhevsky 
et al. 2012), VGG (Simonyan and Zisserman 2014), GoogLeNet (Szegedy et al. 
2015) and ResNet (He et al. 2016) due to their superior performance on different 
benchmarks for object recognition tasks. A typical trend with the evolution of these 
architectures is that the networks have gotten deeper, e.g., ResNet is about 19, 8 and 7 
times deeper than AlexNet, VGGNet, and GoogLeNet respectively. While the increasing 
depth improves feature representation and prediction performance, it also increases 
complexity, making it difficult to optimize and even becomes prone to overfitting. Further 
ther, the increasing number of layers (i.e., network depth) leads to vanishing gradient 
problems as a result of a large number of multiplication operations. Hence, many 
teams chose the DenseNet (Iandola et al. 2014) which connects each layer to every
other layer in a feed-forward fashion, reducing the number of training parameters and alleviates the vanishing gradient problem. DenseNet exhibits $$$\ell (\ell + 1)/2$$ connections in $$$\ell$$ layer network, instead of only $$$\ell$$, as in the networks mentioned above. This enables feature reuse throughout the network that leads to more compact internal representations and in turn, enhances its prediction accuracy. Another opted approach, Deep Layer Aggregation (DLA) structures [Yu et al., 2017], extends the “shallow” skip connections in DenseNet to incorporate more depth and sharing of the features. DLA uses two structures – iterative deep aggregation (IDA) and hierarchical deep aggregation (HDA) that iteratively and hierarchically fuse the feature hierarchies (i.e. semantic and spatial) to make networks work with better accuracy and fewer parameters. Recent Fully Convolutional Network (FCN) [Long et al., 2015] adapt and extend deep classification architectures (VGG and GoogLeNet) into fully convolutional networks and transfer their learned representations by fine-tuning to the segmentation task. It defines a skip architecture that combines semantic information from a deep, coarse layer with appearance information from a shallow, fine layer to produce accurate and detailed segmentations.

For the lesion segmentation task, most of the participating teams exploit U-Net architecture [Ronneberger et al., 2015]. The main idea in U-Net architecture is to supplement the usual contracting network through a symmetric expansive path by addition of successive layers, where upsampling (via deconvolution) is performed instead of pooling operation. The upsampling part consists of large number of feature channels, that allow the network to propagate context information to higher resolution layers. The high resolution features from the contracting path are merged with the upsampled output and fed to soft-max classifier for pixel-wise classification. This network works with very few training images and enables the seamless segmentation of high resolution images by means of an overlap-tile strategy. Other similar architecture SegNet [Badri-narayanan et al., 2015] was opted by a team, it consists of an encoder and decoder network, where the encoder network is topologically identical to the CONV layers in VGG16 and in which FC layer is replaced by a softmax layer. Whereas, the decoder network comprises a hierarchy of decoders, one corresponding to each encoder. The decoder uses max-pooling indices for upsampling its encoder input to produce a sparse
feature maps. Later, it convolves the sparse feature maps with a trainable filter bank to
densify them. At last, the decoder output is fed to a soft-max classifier for generation
of segmentation map. One team choose Mask R-CNN (He et al. [2017], a technique
primarily based on a Region Proposal Network (RPN) that shares convolutional fea-
tures of entire image with the detection network, thus enabling region proposals to
localize and further segments normal and abnormal structures in the retina. RPN is a
fully convolutional network that contributes in concurrently predicting object bounds
and “objectness” scores at each position.

Following subsections present the solutions designed by participating teams with
respect to three sub-challenges. Table 6 summarizes the data augmentation, normaliza-
tion and preprocessing tasks performed by each team.

Table 6. Summary of data augmentation, normalization and pre-processing in the competing solutions. Where, RF, RR, RS, RT, RC represent random flip, rotation, scaling, translation and crop respectively.

<table>
<thead>
<tr>
<th>Task</th>
<th>Team Name</th>
<th>Data Augmentation</th>
<th>Data Normalization</th>
<th>Data Preprocessing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-challenge 1</td>
<td>VRT</td>
<td>✓ ✓ ✓ ✓ ✓</td>
<td>shear</td>
<td>FOV cropping,</td>
</tr>
<tr>
<td></td>
<td>iFLYTEK</td>
<td>✓ ✓ ✓ ✓ ✓</td>
<td>×</td>
<td>division by 255</td>
</tr>
<tr>
<td></td>
<td>PATech</td>
<td>✓ ✓ ✓ ✓ ✓</td>
<td>×</td>
<td>mean subtraction</td>
</tr>
<tr>
<td></td>
<td>SDNU</td>
<td>✓ ✓ ✓ ✓ ✓</td>
<td>×</td>
<td>lesion patch</td>
</tr>
<tr>
<td></td>
<td>SOONER</td>
<td>✓ ✓ ✓ ✓ ✓</td>
<td>×</td>
<td>extraction</td>
</tr>
<tr>
<td></td>
<td>LzyUNCC</td>
<td>✓ ✓ ✓ ✓ ✓</td>
<td>×</td>
<td>mean subtraction,</td>
</tr>
<tr>
<td></td>
<td>SABIST</td>
<td>✓ ✓ ✓ ✓ ✓</td>
<td>×</td>
<td>lesion patch</td>
</tr>
<tr>
<td>Sub-challenge 2</td>
<td>LzyUNCC</td>
<td>✓ ✓ ✓ ✓ ✓</td>
<td>color, stochastic</td>
<td>FOV cropping, image</td>
</tr>
<tr>
<td></td>
<td>VRT</td>
<td>✓ ✓ ✓ ✓ ✓</td>
<td>✓</td>
<td>and photo-metric</td>
</tr>
<tr>
<td></td>
<td>Mammoth</td>
<td>✓ ✓ ✓ ✓ ✓</td>
<td>✓</td>
<td>enhancement</td>
</tr>
<tr>
<td></td>
<td>AVASAVA</td>
<td>✓ ✓ ✓ ✓ ✓</td>
<td>✓</td>
<td>morphological</td>
</tr>
<tr>
<td></td>
<td>HarangiM1</td>
<td>✓ ✓ ✓ ✓ ✓</td>
<td>✓</td>
<td>opening and</td>
</tr>
<tr>
<td></td>
<td>HarangiM2</td>
<td>✓ ✓ ✓ ✓ ✓</td>
<td>✓</td>
<td>closing</td>
</tr>
<tr>
<td>Sub-challenge 3</td>
<td>DeepDR</td>
<td>✓ ✓ ✓ ✓ ✓</td>
<td>OD, fovea region</td>
<td>FOV cropping,</td>
</tr>
<tr>
<td></td>
<td>VRT</td>
<td>✓ ✓ ✓ ✓ ✓</td>
<td>shear and</td>
<td>mean subtraction</td>
</tr>
<tr>
<td></td>
<td>ZIU-BI SGEX</td>
<td>✓ ✓ ✓ ✓ ✓</td>
<td>cropped OD</td>
<td>FOV cropping,</td>
</tr>
<tr>
<td></td>
<td>SDNU</td>
<td>✓ ✓ ✓ ✓ ✓</td>
<td>×</td>
<td>contrast adjustment</td>
</tr>
<tr>
<td></td>
<td>HTKU KLIV</td>
<td>✓ ✓ ✓ ✓ ✓</td>
<td>✓</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>CBER</td>
<td>✓ ✓ ✓ ✓ ✓</td>
<td>✓</td>
<td>-</td>
</tr>
</tbody>
</table>

1 Reference: Krizhevsky et al. [2012]
2 Reference: Howard [2013]
A. Sub-challenge – I: Lesion Segmentation

For a given image, this task seeks to get the probability of a pixel being a lesion (either MA, HE, EX or SE). Although different retinal lesions have distinct local features, for instance, MA, HE, EX, SE have different shape, color and distribution characteristics, these lesions share similar global features. Hence, majority of the participating teams built a general framework that would be suitable for segmentation of different lesions, summarized as follows:

A.1. VRT (Jaemin Son et al.)

Son et al. modified U-Net in such a way that the upsampling layers have the same number of feature maps with the layers concatenated, based on the motivation that features in initial layers and upsampled layers are equally important to the segmentation, thus should have the same number of feature maps. Additionally, they adjusted the number of max-pooling so that radius of the largest lesion spans a pixel in the most coarse layer. In case of EX and HE, max-pooling is done 6 times, whereas for SE and MA it is done 4 times and twice. Further, for dealing with MA’s, they used inverse pixel shuffling to convert a $1280 \times 1280 \times 3$ pixels image to $640 \times 640 \times 12$ for network input and pixel shuffling (Shi et al., 2016) to convert $640 \times 640 \times 4$ segmentation map into $1280 \times 1280 \times 1$ pixels. Later, the pairs of a normalized fundus image and reference ground truths were fed to the network to generate segmentation result in range $[0,1]$. They used weighted binary cross entropy (Murphy, 2012) as loss function given by

$$L = \frac{1}{N} \sum_{n=1}^{N} \left[ -\alpha G^n \log O^n - (1 - G^n) \log(1 - O^n) \right]$$

(3)

where $N$ denotes the number of the pairs in a batch, $G^n$ and $O^n$ represent true segmentation and predicted segmentation for $n^{th}$ image. The value of $\alpha$ was determined as follows:

$$\alpha = \frac{B^n_0}{\gamma F^n_1}$$

(4)

where $B^n_0$ and $F^n_1$ denote the number of background and foreground pixels in the $n^{th}$ image. Since background overwhelms foreground in the lesion segmentation, this
loss function was designed to penalize false negatives in order to boost sensitivity, an important factor in detecting lesions. Also, \( \gamma \) was left as a hyper-parameter and chosen out of \( \{0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 256, 512\} \) to yield the highest AU-PR on the validation set. The final selected \( \gamma \) values for different lesions are summarized in Table 7. They trained the network over 300 epochs using Adam optimizer (Kingma and Ba, 2014) with hyper-parameters of \( \beta_1 = 0.5, \beta_2 = 0.999 \) and learning rate of \( 2e^{-4} \) until 250 epochs and \( 2e^{-5} \) until the end. All implementation was done by Keras 2.0.8 with tensorflow backend 1.4.0 using a server with 8 TITAN X (pascal). The source code is available at [https://bitbucket.org/woalsdnd/isbi-2018-fundus-challenge](https://bitbucket.org/woalsdnd/isbi-2018-fundus-challenge).

### Table 7. \( \gamma \) values in Eq. 4

<table>
<thead>
<tr>
<th>EXs</th>
<th>SEs</th>
<th>HEs</th>
<th>MAs</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>512</td>
<td>8</td>
<td>32</td>
</tr>
</tbody>
</table>

Wang et al. proposed a novel cascaded CNN based approach for retinal lesion segmentation with U-Net as a base model. It consists of three stages, the first stage is a coarse segmentation model to get initial segmentation masks, then second stage is a cascade classifier which was designed for false positive reduction, at last a fine segmentation model was used to refine results from the previous stages. First stage model was trained using the patches of size \( 256 \times 256 \) pixels centered on the particular lesion amongst MA, HE or EX and \( 320 \times 320 \) pixels for SE, resulting in the coarse segmentation outcome. Results of previous stage are coarse due to the fact that non-focus regions (non target lesions) were not utilized in the learning process leading to high false positive count. In the second stage, unlike the first segmentation model which used a lesion centered sample from input dataset pool, candidate regions were extracted using probability maps from the previous stage. Here, the input size fed to model for SE was \( 320 \times 320 \times 3 \) pixels, for HE and EX it was \( 256 \times 256 \times 3 \) pixels, and for MA it was modified to \( 80 \times 80 \times 3 \) pixels considering its small appearance. In this step, a candidate region was regarded as a positive sample if its intersection-over-
union with the ground truth was greater than the given threshold (i.e. 0.5). In this way, most trivial non-focus regions were effectively rejected. However, it was identified in the test that a small proportion of false positives still exist, so an additional model was introduced to refine the segmentation results. In the last stage, candidate regions survived from the second stage were utilized as the input patches resulting in more accurate segmentation results. For first and third stage, they used binary cross entropy or dice loss function (multi-model training), whereas, for second stage, they used only binary cross entropy as loss function. The first, second and third stage models were trained for 100, 300 and 100 epochs respectively with momentum of 0.9. In which, the initial learning rate for first and third stage was set 0.1 and is reduced by 10 times every 30 epochs, and for second stage it was set to 0.001 reduced by 10 times every 80 epochs. MXNET platform was used for training the models.

A.3. PATech (Liu Lihong et al.)

Lihong et al. developed a novel patch-based CNN model (as shown in Fig. 4) in which they innovatively combined the DenseNets and dilation block with U-Net to capture more context information and multi-scale features. The model is composed of

![Proposed architecture for lesion segmentation](image)

Fig. 4. Proposed architecture for lesion segmentation
a down-sampling path with 4 Transitions Down (TD), 4 Dilation Block (DL) and an up-sampling path with 4 Transitions Up (TU). To capture multi-scale features, DL (see Fig. 5) is used with dilation rate of 1, 3 and 5 are concatenated for the convolution. The dense block (DB) is constructed by four layers. The idea behind novel combination of dilation convolution is to better deal with the lesions appearing at different scales, where small dilation rate pay closer attention to the characteristics of the tiny lesions, larger dilation rate focus on large lesions. On the other hand, use of DB’s enabled a deeper and more efficient network.

Initially, they extracted regions within FOV from the images and then normalized them to eliminate local contrast differences and uneven illumination. Later, they used small patches $256 \times 256$ pixels at stride of 64 (128 for MA) to generate the training samples (only patches that overlap with the lesion ground truth) followed by data augmentation before feeding to the model. To deal with highly imbalanced spread of data, they designed a loss function that is combination of dice function \cite{Sudre} and 2D cross Entropy as follows:

$$
L = -mean(w_{10} * G * log(O)) + w_{11} * (1 - G) * log(1 - O) + w_2 * dice(G))
$$

where $w_{10}$ and $w_{11}$ are the factor utilized to keep a balance between the positive and negative pixels, and $w_2$ is the factor utilized to control the significance between dice

![Fig. 5. Architecture for dilation block.](image-url)
and cross entropy loss. The values of $w_{10}$, $w_{11}$ and $w_{2}$ were empirically set to 0.7, 0.3 and 0.4 respectively. The models were trained using Adam optimizer with default parameters, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate was set to $2 \times 10^{-4}$, and then divided by 20 in every 20 epochs. This model was implemented with pytorch1.12 and Tesla M60 platform was utilized for training on the centos 7.2 operating system.

A.4. SOONER (Yunzhi Wang et. al.)

Wang et. al. adopted the U-Net architecture for solving the retinal lesion segmentation problem. The network takes a $380 \times 380$ pixels fundus image patch as an input and predict the binary mask of retinal lesion within the $196 \times 196$ pixels central region of the input patch. They pre-processed the fundus images by subtracting the local mean of each color channel and performed random flipping for data augmentation. Batch normalization was utilized to improve training efficiency and all convolution operations adopted ‘valid’ paddings. For training, they followed a three-stage process for each type of lesions (i.e. MA, HE, EX and SE). For the first stage, they extracted positive image patches in the training set according to the given ground truth mask, and randomly extracted negative image patches from fundus images with and without apparent retinopathy. The objective function was the summation of cross entropy loss functions for MA, HE, EX and SE. Adam algorithm was employed to optimize the parameters. In the second stage, they fine-tuned the U-Net using the extracted patches for each lesion type. Subsequently, they applied the optimized U-Net on the fundus images in the training set and extracted false positive patches generated by U-Net. They further fine-tuned the U-Net using the positive image patches together with the false-positive patches (hard negative patches) as a third stage. In the testing phase, they extracted overlapped image patches using a sliding window and fed the patches into the network to get the corresponding probability maps. The initial learning rate was set to $1 \times e^{-4}$ and fixed number of steps was used as a stopping criteria. They implemented the U-Net architecture based on TensorFlow library with a Nvidia GeForce GTX 1080Ti GPU.
A.5. LZYUNCC (Zhongyu Li et al.)

Li et al. developed a method based on FCN by embedding DLA structure for the segmentation of HE’s and SE’s. As the lesions are located dispersively and irregularly, embedding of DLA structure with FCN enables better aggregation of semantic and spatial information from local and global levels, providing a boost in recognizing their presence. They used retinal images with pixel-level ground truth annotations from both IDRiD and E-Ophtha database. They first adopted a series of methods for data preprocessing and augmentation. Subsequently, considering the correlation between EX’s and SE’s, they first trained an initial model for the segmentation of EX. They chose a smaller model, i.e., DLA-34 to train the segmentation network with binary cross-entropy as a loss function. At last, the trained deep model was fine-tuned for the segmentation of SE. While the model training of EX segmentation, a trade-off parameter (penalty) was assigned in the loss function to control the weights of foreground pixels, and tried different penalty values from 1 to 16 during the model training. At last, these segmentation results were fused to adaptively compute the best performance. They adopted the original DLA cityscapes segmentation experimental settings (Yu et al., 2017) and trained the model for 100 epochs with batch size 4, where the poly learning rate was \((1 - \frac{epoch-1}{total\text{epoch}})^{0.9}\) with momentum of 0.9. The initial learning rate was set to 0.01.

A.6. SAIHST (Yoon Ho Choi et al.)

Choi et al. proposed a model for segmentation of EX based on U-net, in which the convolution layers of the encoder path are replaced with dense blocks. Whereas, the decoder path of their model was kept identical to that of general U-net. They built the dense block with growth factor of 12 and 3 × 3 convolution layers, batch normalization, and ReLU activation. The last layer generates pixel-level prediction maps for EXs through the sigmoid activation function. For training, they utilized only green channel of fundus image and enhanced it using Contrast Limited Adaptive Histogram Equalization (CLAHE). Later, each image was padded to a size of 4352 × 3072 pixels and cropped into 204 patches of 512 × 512 pixels. These patches are further augmented and used for training. The losses were calculated by the binary cross-entropy. The model was trained for 20 epochs with a mini-batch size of 10 and they used Adam
optimizer with an initial learning rate of $2e^{-4}$, $\beta_1$ of 0.9 and $\beta_2$ of 0.999. The model was programmed in Keras 2.1.4 served with Tensorflow 1.3.0 backend.

A.7. SDNU (Xiaodan Sui et al.)

Sui et al. proposed a method based on Mask R-CNN structure to segment lesions from the fundus image. They adopted implementation of Mask R-CNN from [Abdulla, 2017] for solving the problem. This method could detect different objects while simultaneously generating instance segmentation mask.

Network training precedes the data augmentation process and binary cross entropy was used as a loss function. The initial learning rate was set to 0.02 with momentum of 0.9. They chose ResNet-101 as a backbone. They implemented algorithm in Keras with Tensorflow as backend and processed on 8 NVIDIA TITAN Xp GPUs. The experiment environment was built under Ubuntu 16.06.

B. Sub-challenge – 2: Disease Grading

For a given image, this task seeks to get a solution to produce a severity grade of the diseases i.e. DR (5 class problem) and DME (3 class problem). Summary of participating solutions is as follows:

B.1. LZYUNCC (Zhongyu Li et al.)

Li et al. developed method based on the ResNet by embedding DLA structure for the automated grading of DR and DME. For this work they used IDRiD and Kaggle dataset. Initially, for the given training images, they perform data preprocessing and data augmentation. Subsequently, based on the designed ResNet with DLA structure, initial models are trained using 35,000 retinal images from the Kaggle dataset. Later, they fine-tuned the model using the IDRiD dataset through 5 fold cross validation technique. Finally, the 5 outputs are ensembled together as the final grades for input images. It is important to note that networks for the grading of DR and DME were trained separately. The training was performed by Stochastic Gradient Descent (SGD) with a mini-batch size of 64, while the learning rate starts from 0.001 and is then divided by 10 every 20 epochs, for 30 epochs in total. The other hyper-parameters are fixed to the settings of original DLA ImageNet classification [Yu et al., 2017].
B.2. VRT (Jaemin Son et al.)

Son et al. used network (Son et al., 2018) for DR grading. Kaggle dataset was initially used to pre-train the network and then the model was fine-tuned with the IDRiD data. Penultimate layer was Global Average Pooled (GAP) and connected with FC layer. The entire output is a single value from which L2 loss was calculated against the true label. SGD was used with nesterov momentum of 0.9 as optimizer. Learning rate was set to $10^{-3}$. The model was trained for 100 epochs. Fundus image was normalized in range [0, 1] and the mean was subtracted channel-wise. For grading of DME, the segmented EXs (using the segmentation network proposed in sub-challenge – 1), localized fovea and segmented OD (using the segmentation network proposed in sub-challenge – 3) were utilized for making final decision. With these information, semi-major axis of the segmented OD ($r$) was estimated. Further, the fundus image is divided into three regions as macular region: $\|x - c\| < r$, near macular region: $r < \|x - c\| < 2r$ and remaining region: $2r < \|x - c\|$. where $x$ denotes a point in the image.

Furthermore, several features such as sum of intensity for segmented EX, the number of pixels above the threshold (178 in the [0, 255] scale), the number of pixels for the smallest and largest blob, the mean pixel numbers of blobs are extracted for each area, and binary flag that indicates whether the OD is segmented. Now, features with high importance were selected among numerous features in the initial training due to gradient boosting (for instance, XGBoost) was likely to overfit when provided with overly redundant features. Messidor dataset was added to the given data and out of which 10% of images were left as validation set. Sets of hyper-parameters are searched by grid-search. The combination of hyper parameters that yielded the highest accuracy in validation set was min child eight: 2, subsample: 0.2, colsample by tree: 0.2, $\lambda$: 9.0, $\alpha$: 1.0, and depth: 6. Other hyper-parameters are set to default values. All implementations were done by pytorch v0.4.1 using a server with 8 TITAN X (pascal). The source code is available at https://bitbucket.org/woalsdnd/isbi-2018-fundus-challenge.
B.3. Mammoth (Junyan Wu et al.)

Wu et al. proposed an unified framework that combines deep feature extractor and statistical feature blending to automatically predict the DR and DME severity scores. For DME, they used DenseNet to directly predict severity score. Whereas for DR, Kaggle training dataset was used to pre-train the DenseNet model through a dynamic sampling mechanism to balance the training instances and later fine tuned using the IDRiD dataset. Initially, the background of all images was cropped and resized to $512 \times 512$ pixels. Later, morphological opening and closing are utilized to preserve bright and dark regions. For instance, the morphological opening can erase the EXs and highlight the MAs. Whereas, the closing operation can remove MAs and preserve EXs. These operations can be used to denoise specific levels of classifications, for example, the risk of DME only depends on the location of the EXs. Further, several standard data augmentation methods (as shown in Table 6) are also employed. Mean Squared Error (MSE) and cross-entropy with five classes were the loss functions employed to train the network and SGD for optimization. The initial learning rate was set to 0.0005 with decrement of 0.1 after every 30 epochs. The initial training was done by 200 epochs and fine tuning by 50 epochs. Afterwards, the last layer was removed before final prediction, and its statistical features were aggregated together into a boosting tree. Specifically, 50 pseudo random augmentations were performed to get 50 outputs from last second FC layer (size of 4096), then the mean and standard deviation of 50 feature vectors for each image were computed, and both vectors were then concatenated together for training in LightGBM. The output from second last layer of fine-tuning experiments were used to train a blending model, strategy adopted from team O’s solution of Kaggle DR challenge. Finally, for the disease grading prediction, gradient boosting tree model was built on combined second last layer from pre-trained network and fine-tuned network.

B.4. HarangiM1 (Balazs Harangi et al.)

Harangi et al. proposed an approach for the classification of retinal images via the fusion of two AlexNet, and GoogLeNet. For this aim, they removed a FC and classification layers and interconnect them by inserting a joint FC layer followed by the
classic softmax/ classification layers for the final prediction. In this way, single network architecture was created which allows to train the member CNNs simultaneously. For each \( I^{(n)} \), let us denote the outputs of the final FC layers of the member CNNs by \( \hat{O}_1^{(n)} \), \( \hat{O}_2^{(n)} \). The FC layer of their ensemble aggregates them via

\[
\hat{O}^{(n)} = A_1 \hat{O}_1^{(n)} + A_2 \hat{O}_2^{(n)}
\]

where the weight matrices \( A_1, A_2 \) were of size \( 5 \times 5 \) and initialized as

\[
A_1 = A_2 = \begin{bmatrix}
1/5 & 0 & 0 & 0 & 0 \\
0 & 1/5 & 0 & 0 & 0 \\
0 & 0 & 1/5 & 0 & 0 \\
0 & 0 & 0 & 1/5 & 0 \\
0 & 0 & 0 & 0 & 1/5
\end{bmatrix}
\]

The last two layers of the ensemble were a softmax and a classification one. Let \( O_{SM}^{(n)} \) be the output of the former layer, the MSE was used for optimization as a loss function:

\[
MSE = \frac{1}{2N} \sum_{n=1}^{N} (\hat{O}_{SM}^{(n)} - O^{(n)})^2
\]
times required on the datasets for DR and DME were 96.6 (189 epochs) and 23.4 (50 epochs) hours respectively. Implementation of this work was done in Matlab 2017b. Training was performed using an NVIDIA TITAN X GPU card with 7 TFlops of single precision performance, 336.5 GB/s of memory bandwidth, 3,072 CUDA cores, and 12 GB memory.

B.5. AVSASVA (Varghese Alex et al.)

Alex et al. used ensembles of pre-trained CNNs (on ImageNet dataset), namely, ResNets and DenseNets for the task of disease grading. For the task of grading of DR, two ensembles of CNNs namely “primary” and “expert” classifiers were used. The primary classifier was trained to classify a fundus image as one of the 4 classes viz; Normal, Mild NPDR, Moderate NPDR or S-(N)-PDR, a class formed by clubbing Severe NPDR and PDR. The expert classifier was trained exclusively on Severe NPDR or PDR images and was utilized to demarcate the input image as one of the aforementioned classes. During inference, each fundus image was resized to a dimension of $256 \times 256$ pixels. For the task of grading of DR in fundus images, they used test time augmentation through the “Ten Crop” function defined in PyTorch. The images were first passed through the primary classifier and then through the expert classifier, only if the image was classified as S-(N)-PDR by primary classifier. The final prediction was achieved by using a majority voting scheme.

For DME grading, two ensembles were trained in a one versus rest approach. Ensemble 1 was trained to classify the input as either “image with no apparent EXs” (Grade 0) or “presence of EXs in image” (Grade 1 & Grade 2), while the Ensemble 2 was trained to classify an image as “Grade 2” DME or not (Grade 0 & Grade 1). During inference, the resized images were fed to both ensembles and the final prediction was obtained by combining the two predictions by utilizing a set of user defined rules. Briefly, the user defined rules were: an image was classified as Grade 0 DME if ensemble 1 and ensemble 2 predict the absence of EXs and the absence of grade 2 DME respectively. A scenario wherein ensemble 2 predicts the presence of grade 2 DME, the images were classified under the category “Grade 2 DME” irrespective of the prediction from ensemble 1. Lastly, images were classified as Grade 1 DME if none
of the above conditions were satisfied.

Both models for DR and DME were initialized with the pretrained weights and the
parameters of networks were optimized by reducing the cross entropy loss with ADAM
as the optimizer. The learning rate was initialized to $10^{-3}$ for DR and $10^{-4}$ for DME.
For DR, the learning rate was reduced by a factor of 10% every instance when the
validation loss failed to drop. Each network was trained for 30 epochs and the model
parameters that yielded the lowest validation loss were used for inference. For DME,
the learning rate was annealed step-wise with step size of 10 and the multiplicative
factor of learning rate decay value of 0.9.

B.6. HarangiM2 (Balazs Harangi et al.)

Harangi et al. combined self-extracted, CNN-based features with traditional, hand-
crafted ones for disease classification. They modified AlexNet to allow the embedding
of handcrafted features via a FC layer. In this way, they created a network architec-
ture that could be trained in the usual way and additionally uses domain knowledge.
They extended the FC layer $FC_{fuse}$ originally containing 4096 neurons of AlexNet by
adding 68-dimensional vector containing handcrafted features. Then, the $4164 \times 5$ (or
$4164 \times 3$ for DME) layer $FC_{class}$ was considered for the DR (or DME) classification
task. In this way, both the final weighing $FC_{class}$ of the handcrafted features were
obtained and the 4096 AlexNet features were trained by back propagation.

To obtain the 68 handcrafted features used by the CNN, they employed one image
level and two lesion specific methods. The amplitude-frequency modulation (AM-
FM) method extracts information from an image by decomposing its green channel at
different scales into AM-FM components (Havlicek, 1996). As a result, a 30-element
feature vector was obtained, which reflects the intensity, geometry and texture of the
structures contained in the image (Agurto et al., 2010). Whereas to extract features
related to the lesions MA and EX, they employed two detector ensembles (Antal and
Hajdu 2012, Nagy et al. 2011), which consist of a set of <preprocessing method (PP),
candidate extractor (CE)> pairs organized into a voting system. Such a <PP, CE> pair
was formed by applying the PP to the retinal image and the CE to its output. This way,
a <PP, CE> pair extracts a set of lesion candidates from the input image, acting like
a single detector algorithm. They used output of these ensembles to obtain 38 features related to the number and size of MA's and EX's. The parameters of the architectures were optimized by SGD algorithm in 85 and 50 epochs for DR and DME respectively. Training times were 83.1 (85 epochs) and 46.2 (50 epochs) hours on the datasets for DR and DME. Implementation of this work was done in Matlab 2017b. Training has been performed using an NVIDIA TITAN X GPU card with 7 TFlops of single precision, 336.5 GB/s of memory bandwidth, 3,072 CUDA cores, and 12 GB memory.

C. Sub-challenge – 3: Optic Disc and Fovea Detection

For a given image, this task seeks to get a solution to localize the OD and Fovea. Further, it seeks to get the probability of pixel being OD (OD segmentation). Summary of approaches is detailed as follows:

C.1. DeepDR (Ling Dai et al.)

Dai et al. proposed a novel deep localization method, which allows coarse-to-fine feature encoding strategy for capturing the global and local structures in fundus images, to simultaneously model the two-task learning problem of the OD and fovea localization. They took advantage of the prior knowledge such as the number of landmarks and their geometric relationship to reliably detect the OD and fovea. Specifically, they first designed a global CNN encoder (with a backbone network of ResNet-50) to localize the OD and fovea centers as a whole by solving a regression task. All max pooling layers were replaced with average pooling layers as compared to the original ResNet architecture, due to the fact that the max pooling could lose some useful pixel-level information for the regression to predict the coordinates. This step was used to simultaneously perform the two detection tasks, because of the geometric relationship between OD and fovea, the performance of multi-task learning is better than single task. The predicted output coordinates of this global CNN encoder component were used for detecting the bounding boxes of the target OD and fovea. Then the current center coordinates are refined through a local encoder (with a backbone network of VGG-16) which only localizes the OD center or fovea center of their related bounding boxes. During training stage, they designed the effective data augmentation scheme to
solve the problem of insufficient training data. In particular, to build the training set of the local encoder, the bounding boxes were randomly selected based on the ground truth, for each object several bounding boxes of different positions and scales were cropped. The local encoder can be reused multiple times to approximate the target coordinates. The local encoder was iterated twice for refining centers comprehensively. All three models were initialized from the pre-trained ImageNet network, and replaced the network’s last FC layer and softmax layer by the center coordinates regressor. The regression loss for the center location was the Euclidean loss. The modified loss function for global and local encoders was \(0.045(L_{OD} + L_{fovea})\) and \(0.045(L_{OD}/L_{fovea})\) respectively. Where \(L_{OD}\) and \(L_{fovea}\) are losses for OD and fovea, and scaling factor was introduced since the original Euclidean distance is too large in practice to converge. The proposed learning model was implemented in Caffe framework and trained using SGD with momentum. The FC layers for center regression were initialized from zero-mean Gaussian distributions with standard deviations 0.01 and 0.001. Biases were initialized to 0. The global encoder was trained for 200 epochs, local encoders (OD and fovea both) for 30 epochs respectively. The batch size for the global encoder was 16, and 64 for the other two local encoders. The learning rate was set as 0.01 and was divided by 10 when the error plateaus.

C.2. VRT (Jaemin Son et al.)

Son et al. proposed an OD segmentation model consisting of U-Net and CNN that takes a vessel image and outputs \(20 \times 20\) activation map whose penultimate layer is concatenated to bottleneck layer of the U-Net. Initially, the original images were cropped (3500 \(\times\) 2848 pixels), padded (3500 \(\times\) 3500 pixels) and then resized (640 \(\times\) 640 pixels). Each image was standardized with its mean and standard deviation (std). When calculating the mean and std, values less than 10 (usually artifacts in the black background) are ignored. Vessel images were prepared with an external network Son et al. (2017). Pixel values in a vessel image range from 0 to 1. It uses external datasets DRIONS-DB (Carmona et al. 2008) and DRIVE (Staal et al. 2004) available with OD and vessel ground truths respectively. For augmentation, the fundus images were affine-transformed and additionally OD was cropped and randomly placed on the image
for random number of times (0 to 5). This augmentation was done to prevent the
network from segmenting OD solely by brightness. Pairs of a fundus image and the
vessel segmentation were provided as input and OD segmentations in the resolution of
640 × 640 and 20 × 20 pixels are given as the ground truth. Binary cross entropy is used
as loss function for both U-Net and vessel network with the loss of $L_{total} = L_{U-Net} +
0.1 \times L_{vessel}$. Total 800 epochs are trained via Adam optimizer and decreasing learning
rate with hyper-parameters of $\beta_1 = 0.5, \beta_2 = 0.999$. The learning rate was $2e^{-4}$ until
400 epochs and $2e^{-5}$ until the end. Weights and biases were initialized with Glorot
initialization method (Glorot and Bengio, 2010).

They also proposed a four branch model in which two branches were dedicated
to prediction of locations for OD and fovea from vessels (vessel branches) and
other two branches aim to predict the locations from both fundus and vessels (main
branches). Similar to OD segmentation, penultimate layers of vessel branches were
depth-concatenated to the main branches. After deriving an activation map that repres-
ts probability of containing the anatomical landmark, hard-coded matrix was mul-
tiplied to yield co-ordinates. Original images were cropped as in the segmentation
task and standardized with the identical method and later augmented by flip and rota-
tion to ease the implementation efforts. Mean absolute error was used as loss function
for both outputs with the loss of $L_{total} = L_{main} + 0.3 \times L_{vessel}$. SGD was used
with nestrov momentum of 0.9 as optimizer. Learning rate was set to $10^{-3}$ from $1^{st}$
to $500^{th}$ epochs and $10^{-4}$ from $501^{th}$ to $1000^{th}$ epochs. All implementation were
done in Keras 2.0.8 with tensorflow backend 1.4.0 using a server with 8 TITAN X

C.3. ZJU-BII-SGEX (Xingzheng Lyu et al.)

Lyu et al. utilized Mask R-CNN to localize and segment OD and fovea simultane-
ously. It scans the image and generates region proposals by 2D bounding boxes. Then
the proposals were classified into different classes and compute a binary mask for each
object. They firstly preprocessed the original retinal image into fixed dimensions as
network input. A feature extractor (ResNet-50) with feature pyramid networks (FPN)
generates feature maps at different scales, which could be used for regions of interest (ROI) extraction. Then a region proposal network (RPN) scans over the feature maps and locates regions that contain objects. Finally, a ROI head network (RHN) is employed to obtain the label, mask, and refined bounding box for each ROI. They also incorporated prior knowledge of retinal image as a post-processing step to improve the model performance. They used IDRiD dataset and two subsets in RIGA dataset (Almazroa et al., 2018) (Messidor and BinRushed, 605 images) with OD mask provided. They applied transfer learning technique to train the model. They firstly trained the RHN network by freezing all the layers of FPN and RPN networks and then fine-tuned all layers. The model was implemented on Tensorflow 1.3 and python 3.4 (source code was modified from Abdulla (2017)). The learning rate started from 0.001 and a momentum of 0.9 was used. The network was trained on one GPU (Tesla K80) with 20 epochs.

C.4. IITkgpKLIV (Oindrila Saha et al.)

Saha et al. used SegNet for segmentation of lesions and OD. OD was added as an additional class in the same problem as lesion segmentation, so that the model could better differentiate EXs and OD which have similar brightness levels. However, in contrast to original SegNet, the final decoder output is fed to a sigmoid layer to produce class probabilities for each pixel independently in 7 channels. Each channel has the same size as input image: 536 × 356 pixels and consists of activations in the range [0,1] where 0 corresponds to background and 1 to the presence of corresponding class. Apart from 5 classes i.e. MA, HE, SE, EX and OD, two additional classes: (i) retinal disk excluding the lesions and OD, and (ii) black background form the 7 channels. Images were downsampled to 536 × 356 pixels, preserving the aspect ratio. Additionally, Drishti-GS (Sivaswamy et al., 2014) dataset was used for data augmentation to account for case of absence of lesions. Further, horizontal, vertical and 180 degree flipped versions of the original images were taken. The network was trained using binary cross entropy loss function and Adam optimizer with learning rate 10^{-3} and \( \beta = 0.9 \). Early stopping of the training based on the validation loss is adopted to prevent overfitting. It was observed that the validation loss started to increase after 200 epochs. One more
softmax layer is introduced after the Sigmoid layer for normalizing the value of a pixel for each class across channels. Segmented output is finally upsampled for each class to 4288 × 2848 pixels. All implementations were done in PyTorch using 2x Intel Xeon E5 2620 v3 processor with GTX TitanX GPU 12 GB RAM and 64 GB System RAM.

C.5. SDNU (Xiaodan Sui et al.)

Sui et al. used Mask R-CNN for solving all tasks in this sub-challenge. Mask R-CNN could realize accurate target detection based on proposed candidate object bounding boxes of a RPN to achieve the objective of OD and Fovea localization. At the same time, it could also get the OD segment at the mask predicting branch. The head architecture of Mask R-CNN (ResNet-101 as a backbone) consists of three parallel branches for classification, bounding-box regression, and predicting mask. By this method, the localization of OD and fovea, and segment the mask of OD could be obtained directly. They retrained the network to get the new weight parameter of the framework. During the training phase, the dataset of this challenge was augmented by flipping, resizing and trained by 10-fold cross-validation. After training 2000 epochs, the last trained model is obtained. They implemented this algorithm in Tensorflow and it is processed on 8 NVIDIA TITAN Xp GPUs. The experiment environment is built under Ubuntu 16.06.

C.6. CBER (Ana Mendonça et al.)

Mendonça et al. proposed handcrafted features based approach for the localization and segmentation tasks in this sub-challenge. Distinct methodologies have been developed for detecting and segmenting these structures, mainly based on color and vascular information. The methodology proposed in the context of this challenge includes three inter-dependent modules. Each module performs a single task: OD localization, OD segmentation or fovea localization. While the modules responsible for the OD localization and segmentation were an improved version of two methods previously published [Mendonca et al., 2013] [Dashtbozorg et al., 2015], the method proposed for fovea localization was completely new. Initially, the module associated with the OD localization receives a fundus image and segments the retinal vasculature. Afterwards, the entropy
of the vessel directions is computed and combined with the image intensities in order to find the OD center coordinates. For OD segmentation, the module responsible for this task uses the position of the OD center for defining the region where the sliding band filter [Pereira et al., 2007; Esteves et al., 2012] is applied. The positions of the support points which give rise to the maximum filter response were found and used for delineating the OD boundary. Since a relation between the fovea-OD distance and the OD diameter was known [Jonas et al., 2015], the module responsible for the fovea localization begins by defining a search region from the OD position and diameter. The fovea center is then assigned to the darkest point inside that region.

6. Evaluation Measures

The performance of each sub-challenge was evaluated based on different evaluation metrics. Following evaluation measures were used for different sub-challenges:

A. Sub-challenge – 1

This sub-challenge evaluates the performance of the algorithms for different lesion segmentation tasks, from the submitted grayscale images, using the available binary masks. As in the lesion segmentation task(s) background overwhelms foreground, a highly imbalanced scenario, the performance of this task was measured using area under precision (a.k.a. Positive Predictive Value (PPV)) recall (a.k.a. Sensitivity (SN)) curve (AUPR) [Saito and Rehmsmeier, 2015].

\[
SN = \frac{True\ Positives}{True\ Positives + False\ Negatives}
\]

(10)

\[
PPV = \frac{True\ Positives}{True\ Positives + False\ Positives}
\]

(11)

The curve was obtained by thresholding the results at 33 equally spaced instances i.e. \([0, 8, 16, \ldots, 256]\) in gray levels or \([0, 0.03125, 0.0625 \ldots, 1]\) in probabilities. The AUPR provides a single-figure measure (a.k.a. mean average precision (mAP)), computed over the Set-B, was used to rank the participating methods. This performance
metric was used for object detection in The PASCAL Visual Object Classes (VOC) Challenge [Everingham et al., 2010]. The AUPR measure is more realistic [Boyd et al., 2013; Saito and Rehmsmeier, 2015] for the lesion segmentation performance over the Area under Receiver Operating Characteristics (ROC).

B. Sub-challenge – 2

Let the expert labels for DR and DME be represented by $DR_G(n)$ and $DME_G(n)$. Whereas, $DR_O(n)$ and $DME_O(n)$ are the predicted results, then correct instance is the case when the expert label for DR and DME matches with the predicted outcomes for both DR and DME. This was done since, even with presence of some exudation that may be categorized as mild DR, its location on the retina is also important governing factor (to check DME) to decide overall grade of disease. For instance, EXs presence in the macular region can affect vision of the patient to greater extent and hence, it should be dealt with priority for referral (that may otherwise be missed or cause delay in treatment with the present convention of only DR grading) in the automated screening systems. Hence, disease grading performance accuracy for this sub-challenge, from the results submitted in CSV format for test images (i.e. $N = 103$), is obtained by algorithm 1 as follows:

**Algorithm 1: Computation of disease grading accuracy**

**Data:** Method Results and Labels with DR and DME Grading

**Result:** Average disease grading accuracy for DR and DME

1. for $n = 1, 2, \ldots, N$
2. Correct = 0;
3. if $(DR_O(n) == DR_G(n)) \text{ and } (DME_O(n) == DME_G(n))$ then
4. Correct = Correct + 1;
5. end
6. end
7. Average Accuracy = $\frac{Correct}{N}$

C. Sub-challenge – 3

For the given retinal image, the objective of sub-challenge – 3 (task - 6 and 7) was to predict the OD and fovea center co-ordinates. The performance of results submitted
in CSV format was evaluated by computing the Euclidean distance (ED) (in pixels) between manual (ground truth) and automatically predicted center location. Lower ED indicates better localization. After determining Euclidean distance for each image in the Set-B, i.e. for 103 images, the average distance representing the whole dataset was computed and used to rank the participating methods.

The optic disc segmentation (task - 8) performance is evaluated using Jaccard index \( J \) \cite{Jaccard}. It represents the proportion of overlapping area between the segmented OD \( O \) and the ground truth \( G \).

\[
J = \frac{|O \cap G|}{|O \cup G|} \quad (12)
\]

Higher \( J \) indicates better segmentation. For the segmented results, images in range \([0, 255]\), it was computed at 10 different equally spaced thresholds \([0, 0.1, \cdots, 0.9]\) and averaged to obtain final score.

7. Results

This section reports and discusses the results of all sub-challenges. Performance of all competing solutions on the Set-B for all eight subtasks are divided into three sub-challenge categories and discussed including their leaderboard rank.

A. Sub-challenge – I

In this section, we present the performance of all competing solutions for the lesion segmentation task. All results received from the participating teams were analyzed using the validation measure given in section 6.A. This measure generated a set of precision-recall curves for each of the different techniques. A total of 22 solutions were evaluated for this sub-challenge (a complete list is available on the challenge website) and ranked using the area under precision-recall curve values. Amongst them, only top-4 teams per lesion segmentation task were invited for the challenge workshop and top-3 teams having overall better performance, the solutions developed by the teams that ranked amongst top three for at least three different lesion segmentation tasks, presented their work at ISBI.
Table 8. Sub-challenge – 1 “Off-site” leaderboard highlighting top 4 teams from each lesion (MAs, HEs, SEs and EXs) segmentation task on the testing dataset. It details the approach followed by respective team and external dataset used for training their model (if any).

<table>
<thead>
<tr>
<th>Lesion</th>
<th>Team Name</th>
<th>AUPR</th>
<th>Approach</th>
<th>Ensemble</th>
<th>Input Size (Pixels)</th>
<th>External Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microaneuys</td>
<td>iFLYTEK</td>
<td>0.5017</td>
<td>Cascaded CNN</td>
<td>✓</td>
<td>320 × 320</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>VRT</td>
<td>0.4951</td>
<td>U-Net</td>
<td>×</td>
<td>1280 × 1280</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>PATech</td>
<td>0.4740</td>
<td>DenseNet+U-Net</td>
<td>✓</td>
<td>256 × 256</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>SDNU</td>
<td>0.4111</td>
<td>Mask R-CNN</td>
<td>×</td>
<td>3584 × 2380</td>
<td>×</td>
</tr>
<tr>
<td>Hemorrhages</td>
<td>VRT</td>
<td>0.6804</td>
<td>U-Net</td>
<td>×</td>
<td>640 × 640</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>PATech</td>
<td>0.6490</td>
<td>DenseNet+U-Net</td>
<td>✓</td>
<td>256 × 256</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>iFLYTEK</td>
<td>0.5588</td>
<td>Cascaded CNN</td>
<td>✓</td>
<td>320 × 320</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>SOONER</td>
<td>0.5395</td>
<td>U-Net</td>
<td>×</td>
<td>380 × 380</td>
<td>×</td>
</tr>
<tr>
<td>Soft Exudates</td>
<td>VRT</td>
<td>0.6995</td>
<td>U-Net</td>
<td>×</td>
<td>640 × 640</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>LzyUNCC-I</td>
<td>0.6607</td>
<td>FCN+DLA</td>
<td>×</td>
<td>1024 × 1024</td>
<td>E-ophtha</td>
</tr>
<tr>
<td></td>
<td>iFLYTEK</td>
<td>0.6588</td>
<td>Cascaded CNN</td>
<td>✓</td>
<td>320 × 320</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>LzyUNCC-II</td>
<td>0.6259</td>
<td>FCN+DLA</td>
<td>×</td>
<td>1024 × 1024</td>
<td>E-ophtha</td>
</tr>
<tr>
<td>Hard Exudates</td>
<td>PATech</td>
<td>0.8850</td>
<td>DenseNet+U-Net</td>
<td>✓</td>
<td>256 × 256</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>iFLYTEK</td>
<td>0.8741</td>
<td>Cascaded CNN</td>
<td>✓</td>
<td>320 × 320</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>SAIHST</td>
<td>0.8582</td>
<td>U-Net</td>
<td>×</td>
<td>512 × 512</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>LzyUNCC-I</td>
<td>0.8202</td>
<td>FCN+DLA</td>
<td>×</td>
<td>1024 × 1024</td>
<td>E-ophtha</td>
</tr>
</tbody>
</table>

Table 8 summarizes the individual performance (Off-site evaluation) of each solution listed in order of their final placement for each subtask. It also contains the various approaches followed and external dataset (if any) used for training the models. The higher the rank for individual task, the more favorable the performance. The top-3 entries according to the individual lesion segmentation task are VRT, iFLYTEK-MIG and PATech. Some sample lesion segmentation results illustrated in Fig. 6 and their corresponding overall evaluation score from Table 8 give a better idea of how the evaluation scores correlate with the quality of the segmentation.

Fig. 7 summarizes the performance of top-4 teams per lesion segmentation task. The different curves represent the performance of the participating methods for various lesions (MAs, HEs, SEs and EXs). Team VRT achieved highest AUPR score for HE and SE segmentation task. Whereas, team PATech and iFLYTEK-MIG obtained best score for EX and MA segmentation task respectively.
Fig. 6. Illustration of lesion segmentation results: (a) sample image and (b) segmentation outcome of top-4 teams (from left to right) (i) MAs, (ii) HEs, (iii) SEs, and (iv) EXs in retinal fundus images. Top row corresponds to ground truths, second row to entry from top performing team, similarly, third, fourth and fifth rows correspond to entries from other three teams respectively. The lesion segmentation entries are colored for better illustration and separation from each type of lesion.
Fig. 7. The AUPR curves for the four top performing individual methods on the test dataset. These curves plot the sensitivity versus the positive predictive values for the different lesions, namely, (a) MAs, (b) HEs, (c) SEs, and (d) EXs.

B. Sub-challenge – 2

This section presents the results achieved (On-site evaluation) by the participating teams for the DR and DME grading task. It is important to note that this task was evaluated for simultaneous grading of DR and DME using the validation algorithm outlined in section 6.B on the test set (Set-B). This algorithm produced an average grading accuracy of joint DR and DME on all images. Table 9 summarizes the result.
of teams for on-site challenge along with the approach followed and the external dataset
used for training the model by respective team.

Table 9. Sub-challenge – 2 “On-site” leaderboard highlighting top 6 teams performance in DR and DME
grading on the testing dataset. It details the approach followed by respective team and external dataset used
for training their model

<table>
<thead>
<tr>
<th>Team Name</th>
<th>Accuracy</th>
<th>Approach</th>
<th>Ensemble</th>
<th>Input Size (Pixels)</th>
<th>External Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>LzyUNCC</td>
<td>0.6311</td>
<td>Resnet + DLA</td>
<td>5</td>
<td>896 × 896</td>
<td>Kaggle</td>
</tr>
<tr>
<td>VRT</td>
<td>0.5534</td>
<td>CNN</td>
<td>10</td>
<td>640 × 640</td>
<td>Kaggle, Messidor</td>
</tr>
<tr>
<td>Mammoth</td>
<td>0.5146</td>
<td>DenseNet</td>
<td>✓</td>
<td>512 × 512</td>
<td>Kaggle</td>
</tr>
<tr>
<td>HarangiM1</td>
<td>0.4757</td>
<td>AlexNet + GoogLeNet</td>
<td>2</td>
<td>224 × 224</td>
<td>Kaggle</td>
</tr>
<tr>
<td>AVSASVA</td>
<td>0.4757</td>
<td>ResNet + DenseNet</td>
<td>DR-8, DME-5</td>
<td>224 × 224</td>
<td>DiaretDB1</td>
</tr>
<tr>
<td>HarangiM2</td>
<td>0.4078</td>
<td>AlexNet + Handcrafted features</td>
<td>2</td>
<td>224 × 224</td>
<td>Kaggle</td>
</tr>
</tbody>
</table>

The top performing solution at the “on-site” challenge was proposed by team LzyUNCC followed by team VRT and team Mammoth. Fig. shows the average accuracy of the competing solutions for individual as well as simultaneous for DR and DME grading task. Teams are observed to perform poorly in the DR grading task that reduced the overall accuracy for simultaneous grading of DR and DME. Major reason seems to be the difficult test set, difficulty in accurately discriminating the DR severity grades.
This section presents the evaluation of “On-site” results for the participating teams in the sub-challenge – 3, for all three subtasks. The results for subtasks of OD and Fovea center localization were evaluated by euclidean distance, whereas for OD segmentation results were evaluated and ranked using Jaccard similarity score as outlined in section 6.C. Results from the on-site evaluations are reported in Table 10 and Table 11 that summarize the results of all participating algorithms for all three subtasks.

Table 10. “On-site” leaderboard highlighting performance of top 5 teams in OD and fovea localization. It highlights the approach followed by respective team and external dataset used for training their model (if any). ED: Euclidean Distance.

<table>
<thead>
<tr>
<th>Localize</th>
<th>Team Name</th>
<th>ED</th>
<th>Rank</th>
<th>Approach</th>
<th>Input Size (Pixels)</th>
<th>External Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optic Disc</td>
<td>DeepDR</td>
<td>21.072</td>
<td>1</td>
<td>ResNet + VGG</td>
<td>224 × 224, 950 × 950</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>VRT</td>
<td>33.538</td>
<td>2</td>
<td>U-Net</td>
<td>640 × 640</td>
<td>DRIVE</td>
</tr>
<tr>
<td></td>
<td>ZJU-BII-SGEX</td>
<td>33.875</td>
<td>3</td>
<td>Mask R-CNN</td>
<td>1024 × 1024</td>
<td>RIGA</td>
</tr>
<tr>
<td></td>
<td>SDNU</td>
<td>36.220</td>
<td>4</td>
<td>Mask R-CNN</td>
<td>1984 × 1318</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>CBER</td>
<td>29.183</td>
<td>-</td>
<td>Handcrafted Features</td>
<td>536 × 356</td>
<td>-</td>
</tr>
<tr>
<td>Fovea</td>
<td>DeepDR</td>
<td>64.492</td>
<td>1</td>
<td>ResNet + VGG</td>
<td>224 × 224, 950 × 950</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>VRT</td>
<td>68.466</td>
<td>2</td>
<td>U-Net</td>
<td>640 × 640</td>
<td>DRIVE</td>
</tr>
<tr>
<td></td>
<td>SDNU</td>
<td>85.400</td>
<td>4</td>
<td>Mask R-CNN</td>
<td>1984 × 1318</td>
<td>RIGA</td>
</tr>
<tr>
<td></td>
<td>ZJU-BII-SGEX</td>
<td>570.133</td>
<td>4</td>
<td>Mask R-CNN</td>
<td>1024 × 1024</td>
<td>RIGA</td>
</tr>
<tr>
<td></td>
<td>CBER</td>
<td>59.751</td>
<td>-</td>
<td>Handcrafted Features</td>
<td>536 × 356</td>
<td>-</td>
</tr>
</tbody>
</table>

The winning methods for the detection task were developed by team DeepDR and team VRT, with DeepDR performing best in both OD and Fovea detection tasks. But the winning entries for OD segmentation task were from teams ZJU-BII-SGEX, VRT and IITKgpKLIV. Some sample OD segmentation results from these teams are illustrated in Fig. 9.

Fig. 10 shows box-plots (McGill et al., 1978) illustrating the range of Euclidean distances from the center of (a) optic disc and (b) fovea as well as (c) spread of Jaccard index for optic disc segmentation.
Table 11. "On-site" leaderboard highlighting performance of top 5 teams in OD segmentation. It details the approach followed by respective team and external dataset used for training their model (if any). J: Jaccard Index.

<table>
<thead>
<tr>
<th>Team Name</th>
<th>J</th>
<th>Rank</th>
<th>Approach</th>
<th>Input Size (Pixels)</th>
<th>External Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZJU-BII-SGEX</td>
<td>0.9338</td>
<td>1</td>
<td>Mask R-CNN</td>
<td>1024 x 1024</td>
<td>RIGA</td>
</tr>
<tr>
<td>VRT</td>
<td>0.9305</td>
<td>2</td>
<td>U-Net</td>
<td>640 x 640</td>
<td>DRIVE, DRIONS-DB</td>
</tr>
<tr>
<td>IITKgpKLIV</td>
<td>0.8572</td>
<td>3</td>
<td>SegNet</td>
<td>536 x 356</td>
<td>Drishti-GS</td>
</tr>
<tr>
<td>SDNU</td>
<td>0.7892</td>
<td>4</td>
<td>Mask R-CNN</td>
<td>1984 x 1318</td>
<td>-</td>
</tr>
<tr>
<td>CBER</td>
<td>0.8912</td>
<td>-</td>
<td>Handcrafted Features</td>
<td>536 x 356</td>
<td>-</td>
</tr>
</tbody>
</table>

Fig. 9. Illustration of OD segmentation results: (a) sample image, (b) optic disc ground truth and (c) segmentation outcome of top-4 teams (from left to right)

8. Discussion and Conclusion

In this paper, we have presented the details of IDRiD challenge with detail information about the data, evaluation metrics, an organization of the challenge, competing solutions and final results for all sub-tasks, i.e., lesion segmentation, disease grading and detection and segmentation of other normal retinal structures. Given the significant number of participating teams (37) and results obtained, we believe this challenge was a success. To the organizational end, efforts have been made in creating a relevant, stimulating and fair competition, capable of advancing the collective knowledge.
in the research community. This section presents a discussion, limitations, and lessons learned from this challenge.

The first sub-challenge was conducted in an off-site mode in which 22 teams participated with their lesion segmentation methods. The results of these methods on the Set-B were evaluated by the organizers and amongst them, top-4 performing methods per lesion segmentation task are included in this paper. The computed AUPR values ranged between 0.4111 (for MAs) and 0.885 (for EXs). The best approach for lesion segmentation used U-net, with data augmentation and the addition of dense block extract the features efficiently, boosting the results significantly. Fig. 11 highlights the performance of top solution for EX that performs significantly well in presence of
normal retinal structures and different challenging circumstances. From the top per-

Fig. 11. Illustration of (a-d) different challenging circumstances for segmentation of EXs, (e-f) segmentation results (probability map) of top-performing team for EXs, (i) enlarged part of Fig. (d), and (j) depicts its performance to be better than (k) the human annotator (The annotator tool had limitation of the markup capability when there is an overlap of multiple types of lesion. In this case, EXs and HE).

forming approaches, it is evident that solving the data imbalance problem improves
the model performance significantly. Since background overwhelms foreground, the
loss during training is more effectively back-propagated than that of foreground that
penalizes false negatives, boosting the sensitivity of lesion segmentation. Architectural
modifications to U-Net-based networks provided widely varying results for the differ-
et types of lesion. For instance, the cascaded CNN approach yielded the best score
for MAs segmentation, as it add modules to reduce false positives. This approach dra-
matically impacts MA segmentation performance due the class imbalance of the task.
Further, Fig. [12] shows that some false positives detected by the participating solutions
are due to noise, predominantly for MA and HE. This indicates that there is still room
for improvement for lesion segmentation tasks with current fundus cameras.

In the on-site disease-grading task six methods were compared and contrasted.
When assessed using the test data set hidden from the participants, the grading ac-
curacy ranged between 0.4078 and 0.6311 as shown in Table [9]. Notably, all teams
except AVASAVA used the external Kaggle DR dataset for pre-training their models.
This dataset contains a large amount of retina images annotated with the disease level, in contrast, team AVASAVA pre-trained their model on ImageNet, a dataset containing natural images and object annotations, effectively showing the network a much smaller number of retina images at training stage, approximately 1% compared to the other teams. This indicates that in the presence of a limited number of labeled data, transfer learning approaches along with the good model pruning could yield comparable and competitive results. However, while the models do determine the variability of the performance, the number, type and quality of training data is a crucial factor for a fair comparison of competing solutions. There is still work needed on simultaneous grading of DR and DME as the reported results do not yet reach the performance needed for a clinically viable automatic screening. Considering the misclassified instances in the confusion matrices in Table 12, along with the lesion information, it is essential to give attention towards characterization of intra-retinal micro-vascular abnormalities (IRMA’s) and venous beading for improvement in the overall grading results.

In the sub-challenge – 3, another on-site challenge, four teams were evaluated for the task of OD/fovea localization and OD segmentation. For the task of OD localization, the Euclidean Distance varied between 21.072 and 36.22 (lower values indicate better performance). However, for Fovea localization task the same performance metric ranged between 64.492 and 570.133. This massive variation is due to outliers, e.g.
team ZJU-BII-SGEX had 23 outliers whose Euclidean Distance exceeded 700. In the OD segmentation task, the average Jaccard similarity index score amongst the participants ranged between 0.7892 and 0.9338. The top-performing solutions developed by DeepDR and VRT leveraged prior clinical knowledge, such as the number of landmarks and their geometric relationship to detect another retinal landmark. It is also observed that data augmentation and ensemble of models yield substantial improvements in terms of accuracy. Considering the clinical significance of OD diameter while DME severity grading, we further compute the average OD diameter (in pixels) for each image of test set. Fig. 13 illustrates the performance of each participating team.

**Table 12. Confusion matrix of retinal images predicted by top performing solution for DR (5 class) and DME (3 class).**

<table>
<thead>
<tr>
<th>Actual</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>30</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>13</td>
<td>3</td>
<td>50</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>41</td>
</tr>
</tbody>
</table>

Fig. 13. Illustration of average OD diameter result of all 5 teams for each image of the testing dataset. (Here the legends G, ZJU and IITkgp represent Groundtruth, ZJU-BII-SGEX and IITKgpKLIV respectively (compressed to appear clearly in single column format, appears full in double column format).)

with respect to the ground truth, most methods show a similar pattern. The average di-
ameter of OD ground truth is 516.61 pixels whereas, this corresponding values for for
the results of solutions developed by the teams ZJU-BII-SGEX, VRT, IITKgpKLIV, 
CBER and SDNU are 514.25, 519.21, 513.48, 508.04 and 460.19 pixels respectively. 
Team CBER submitted their after the competition and they were not included in the 
leaderboard.

As expected, we found that image resolution is a vital factor for the model perfor-
mance, especially for the task of segmentation of small objects such as MAs or EXs. 
In fact, the top performing approaches process the images patch-wise, which allow 
models to have a local high resolution image view or directly with the high resolu-
tion image as a whole. This is essential as MAs or small EXs lesions span very few 
pixels in some cases, and reducing the original image size would prevent an accurate 
segmentation. Similarly, image resolution plays a very important role for the disease 
classification task (see Table 9), the most likely reason is that the presence of the dis-
ease is determined by the presence of lesions in the image, including the small ones 
that might be invisible at low resolution. This is corroborated by the confusion matri-
ces in Table 12 which show misclassified instances in DR (particularly, grade 1 and 2) 
as well as DME (5 images each belonging to grade 1 and 2 are predicted as grade 0). 
For the localization tasks, all participants were asked to identify retinal structures with 
coordinates at full image resolution. Most of them performed these tasks by scaling the 
image to smaller size and then converted their predictions in the original image space. 
The results indicate that the input image resolution has limited effect on the results of 
the localization problem. For instance, in case of OD localization, the top performing 
team utilized two image resolutions, one (224 × 224 pixels) for approximate location 
prediction and other (cropped ROIs 950 × 950 pixels) for refining that estimate. Sim-
ilarly, teams CBER and VRT resized the image to 536 × 356 pixels and 640 × 640 
pixels respectively to get an approximate center location whereas, the team SDNU uti-
lized the input size of 1984 × 1318 pixels. Considering the OD average diameter of 
approximately 516 pixels, the deflection of result for about 10 to 15 pixels by other 
approaches, utilizing different input resolutions, as compared to the top performing so-
lution is very less. This is because the retinal structures to be identified, OD and fovea, 
are very unlikely to disappear due to a reduction of image resolution and they have
This challenge provides data collected in the routine clinical practice and the acquisition protocol was consistent for all images. The data was acquired after pupil dilation with the same camera at the same resolution, ensuring a consistent quality. This dataset did not include non-gradable images and images with substantial disagreement amongst the expert annotators. Even after these efforts to provide the best possible data, the annotation process is still inherently subjective and the annotator judgement is a limiting factor for the method performance which are mostly trained and evaluated in a supervised manner. While we believe that data challenges like ours foster “methodology diversity”, the majority of competing solutions used deep convolutional networks. These approaches are comparably easier to implement than approaches based on feature engineering and do generalize well to multiple medical imaging domains dramatically reduces the need for specialized task knowledge. Notably, amongst the competing solutions in this challenge that utilized deep learning approach along with the task-relevant subject knowledge have demonstrated superior performance. However, it seems there might be some impact of challenge duration, apart from the number of submissions, on the quality of developed solutions. Considering the time span from data availability to deadline of results submission, about one and a half month, was considerably tight for managing all tasks at the same time. For the team VRT who had been working on analyzing fundus images for more than a year when participated in the competition that attempting all tasks were possible, still it was challenging for them to commit all the tasks. However, it would be highly challenging for a newcomer to succeed in multiple tasks. In that sense, the competition period was not sufficient for perfecting all tasks. However, it would be enough for a competent participant, e.g. new entrants in the field as team SAIHST, to finish one task if the participant can focus on the competition completely. Also, in this challenge, the results were evaluated all at once after the result submission deadline. A continuous on-line assessment of participating solutions would have facilitated the submission procedure by providing real-time feedback to the team’s performance. This would have enabled a maximum number of submissions during the challenge period, probably boosting the final count of submissions. However, this would have introduced a risk of overfitting the test data.
by continuous submissions based on the system’s performance on the test set.

This challenge led to the development of a variety of new robust solutions for lesion segmentation, detection, and segmentation of retinal landmarks and disease severity grading. Despite the complexity of the tasks, less than one-and-a-half month time for development, it received a very positive response, and the top performing solutions were able to achieve results close to the human annotators. Still, there is room for improvement, especially in the lesion segmentation and disease-grading tasks. Though the competition is now completed, the dataset has been made publicly available for research purposes to attract newcomers to the problem and to encourage the development of novel solutions to meet current and future clinical standards.

Acknowledgments

This work is sponsored by the Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded (M.S.), INDIA. The authors would like to thank the following people for their help in various aspects of organizing the ISBI-2018 Diabetic Retinopathy Segmentation and Grading Challenge: Prof. Emanuele Trucco (University of Dundee, Scotland) and Tom MacGillivray (University of Edinburgh, Scotland), Ravi Kamble (SGGS Institute of Engineering and Technology, Nanded), Prof. Vivek Sahasrabuddhe (Government Medical College, Nanded) and Désiré Sidibé (Université de Bourgogne, France). We would also like to thank Prof. Jorge Cuadros, University of California, Berkeley (Organizer of Kaggle Diabetic Retinopathy challenge) for his kind permission for reporting the results of the models trained on their dataset.

VRT. : This study was supported by the Research Grant for Intelligence Information Service Expansion Project, which is funded by National IT Industry Promotion Agency (NIPA-C0202-17-1045) in South Korea.

DeepDR. : This work was supported in part by the National Natural Science Foundation of China under Grant Grant 61872241, Grant 61572316, in part by the National Key Research and Development Program of China under Grant 2016YFC1300302 and Grant 2017YFE0104000, in part by the Science and Technology Commission of Shanghai Municipality under Grant 16DZ0501100 and Grant 17411952600.
HarangiM1-M2. Research was supported in part by the project EFOP-3.6.2-16-2017-00015 supported by the European Union and the State of Hungary, co-financed by the European Social Fund.

ZJU-BII-SGEX. This work is supported by Beijing Shanggong Medical Technology Co., Ltd., which provided ocular healthcare solutions in China. Many thanks to the labeled images from Image Annotation Group of Beijing Shanggong Medical Technology.

Team CBER (A.M. Mendonça, T. Melo, T. Araújo and A. Campilho) is financed by the ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme, and by National Funds through the FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project CMUP-ERI/TIC/0028/2014. Teresa Araújo is funded by the FCT grant SFRH/BD/122365/2016.

Appendix A. Comparison of Publicly Available Retinal Image Databases

Table A.1 and Table A.2 provides the summary of technical specifications and available ground truths in several existing datasets and the IDRiD dataset.

References


Table A.1. Summary of technical specifications and hardware used in different databases

<table>
<thead>
<tr>
<th>Name of Database</th>
<th>Number of Images</th>
<th>Image Size(s)</th>
<th>FOV</th>
<th>Camera</th>
<th>Technical Details</th>
<th>NMY</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARIA</td>
<td>212</td>
<td>768×576</td>
<td>50</td>
<td>Zeiss FF450+</td>
<td>✓</td>
<td>✓</td>
<td>TIFF</td>
</tr>
<tr>
<td>DIARETDB</td>
<td>130+89</td>
<td>1500×1152</td>
<td>50</td>
<td>Zeiss FF450+</td>
<td>✓</td>
<td>✓</td>
<td>PNG</td>
</tr>
<tr>
<td>DRIVE</td>
<td>40</td>
<td>768×584</td>
<td>45</td>
<td>Canon CR5</td>
<td>✓</td>
<td>✓</td>
<td>JPEG</td>
</tr>
<tr>
<td>E-Optha</td>
<td>47EX+35H 148MA+233H</td>
<td>1440×960 - 2048×1360 (4)</td>
<td>45</td>
<td>Canon CR - DG1 &amp; Topcon TRC - NW6</td>
<td>✓</td>
<td>✓</td>
<td>JPEG</td>
</tr>
<tr>
<td>HEIMED</td>
<td>169</td>
<td>2196×1958</td>
<td>45</td>
<td>Zeiss Visucam PRO</td>
<td>✓</td>
<td>✓</td>
<td>JPEG</td>
</tr>
<tr>
<td>Kaggle</td>
<td>88,702</td>
<td>433×289 - 3888×2592</td>
<td>Varying</td>
<td>Any camera (EyePACS Platform)</td>
<td>-</td>
<td>✓</td>
<td>TIFF</td>
</tr>
<tr>
<td>MESSIDOR</td>
<td>800 MY+ 400 NMY+ 1756</td>
<td>1440×960, 2240×1488, 2304×1536</td>
<td>45</td>
<td>3CCD/Topcon TRC NW6</td>
<td>✓</td>
<td>✓</td>
<td>TIFF</td>
</tr>
<tr>
<td>ROC</td>
<td>100</td>
<td>768×576, 1058×1061, 1389×1383</td>
<td>45</td>
<td>Topcon NW100 &amp; NW200</td>
<td>✓</td>
<td>✓</td>
<td>JPEG</td>
</tr>
<tr>
<td>STARE</td>
<td>397</td>
<td>605×700</td>
<td>35</td>
<td>Topcon TRV - 50</td>
<td>✓</td>
<td>✓</td>
<td>PPM</td>
</tr>
<tr>
<td>IDRiD</td>
<td>516 (81 with LA)</td>
<td>4288×2848</td>
<td>50</td>
<td>Kowa VX - 10α</td>
<td>✓</td>
<td>✓</td>
<td>JPG</td>
</tr>
</tbody>
</table>


Table A.2. Comparison of different databases with the IDRiD database

<table>
<thead>
<tr>
<th>Name of Database</th>
<th>Normal Fundus Structures</th>
<th>Abnormalities</th>
<th>Multiple Experts</th>
<th>Disease Grading</th>
<th>Diabetic Macular Edema</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OD VS FA MA HE EX SE Yes/No #</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARIA</td>
<td>✓ ✓ ✓ ✓ × × × ✓ ✓ ✓ ✓ ✓ ✓</td>
<td>2 ✓ × ×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIARETDB</td>
<td>× × × ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓</td>
<td>4 × × ×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DRIVE</td>
<td>✓ ✓ × × × ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓</td>
<td>3 × ×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-Optha</td>
<td>× × × ✓ × ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓</td>
<td>2 ✓ ×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HEIMED</td>
<td>× × × × × × × × × × × ✓ ✓ ✓</td>
<td>1 ✓ ✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kaggle</td>
<td>× × × × × × × × × × × ✓ ✓ ✓</td>
<td>2 ✓ ×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MESSIDOR</td>
<td>× × × × × × × × × × × ✓ ✓ ✓</td>
<td>1 ✓ ✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROC</td>
<td>× × × ✓ × × × ✓ × × × ✓ ✓ ✓</td>
<td>4 × ×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STARE</td>
<td>✓ ✓ ✓ ✓ × × × × × × × × × ✓</td>
<td>2 × ×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IDRiD</td>
<td>✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓</td>
<td>2 ✓ ✓</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

OD - Optic Disc, MC - Macula, VS - Vessels, FA - Fovea, MA - Microaneurysms, HE - Hemorrhage, EX - Hard Exudate, SE - Soft Exudate, # - Number of Experts

Systems 32 (6), 481–488.


URL [http://diabetesatlas.org/resources/2017-atlas.html](http://diabetesatlas.org/resources/2017-atlas.html)


Decencière, E., Zhang, X., Cazuguel, G., Laÿ, B., Cochener, B., Trone, C., Gain, P.,
distributed image database: the messidor database. Image Analysis and Stereology
33 (3), 231–234.
Dhara, A. K., Mukhopadhyay, S., Bency, M. J., Rangayyan, R. M., Bansal, R., Gupta,
A., 2015. Development of a screening tool for staging of diabetic retinopathy in
International Society for Optics and Photonics, p. 94140H.
Dobbin, K. K., Simon, R. M., 2011. Optimally splitting cases for training and testing
Esteves, T., Quelhas, P., Mendonça, A. M., Campilho, A., 2012. Gradient convergence
filters and a phase congruency approach for in vivo cell nuclei detection. Machine
pascal visual object classes (voc) challenge. International journal of computer vision
88 (2), 303–338.
Farnell, D. J., Hatfield, F., Knox, P., Reakes, M., Spencer, S., Parry, D., Harding, S.,
2008. Enhancement of blood vessels in digital fundus photographs via the application
of multiscale line operators. Journal of the Franklin institute 345 (7), 748–765.
269 (10), 1290–1291.
Figueiredo, I. N., Kumar, S., Oliveira, C. M., Ramos, J. D., Engquist, B., 2015. Auto-
mated lesion detectors in retinal fundus images. Computers in biology and medicine
66, 47–65.


Hatanaka, Y., Nakagawa, T., Hayashi, Y., Kakogawa, M., Sawada, A., Kawase, K., Hara, T., Fujita, H., 2008. Improvement of automatic hemorrhage detection meth-


Khojasteh, P., Júnior, L. A. P., Carvalho, T., Rezende, E., Aliahmad, B., Papa, J. P.,
features. Computers in biology and medicine.

for medical imaging. Precision and Future Medicine 2 (2), 37–52.


convolutional neural networks. In: Advances in neural information processing sys-

Lam, C., Yu, C., Huang, L., Rubin, D., 2018. Retinal lesion detection with deep learn-
ing using image patches. Investigative ophthalmology & visual science 59 (1), 590–
596.

Li, H., Chutatape, O., 2004. Automated feature extraction in color retinal images by a
model based approach. IEEE Transactions on biomedical engineering 51 (2), 246–
254.

Lin, S., Ramulu, P., Lamoureux, E. L., Sabanayagam, C., 2016. Addressing risk fac-
tors, screening, and preventative treatment for diabetic retinopathy in developing

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van

segmentation. In: Proceedings of the IEEE conference on computer vision and


URL http://dx.doi.org/10.21227/H25W98


Saito, T., Rehmsmeier, M., 2015. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. PloS one 10 (3), e0118432.


75


IDRiD: Diabetic Retinopathy - Segmentation and Grading Challenge
Highlights

• Outlines the setup of challenge on “Diabetic Retinopathy – Segmentation and Grading” held at ISBI-2018.
• Describes the dataset used, evaluation criteria and results of top performing participating solutions.
• Presents the details of various handcrafted feature and deep learning based participating approaches.
• Discusses the lessons learnt from the analysis of the methods submitted to this challenge.
Response to Reviewers Comments

Manuscript Reference: #MEDIA-D-19-00049

Manuscript title: IDRiD: Diabetic Retinopathy – Segmentation and Grading Challenge

We would like to thank all the reviewers and editor-in-chief for their careful reading of the manuscript and thoughtful comments which resulted in improving overall quality of the manuscript. The paper has now been duly revised in accordance with these comments. A point by point response to the reviewers follows.

Comments to the Author

Reviewer #1:

Comments:

Manuscript Rating Question(s): Scale [1-5]
The paper is of enough importance to warrant publication in MedIA 4
The paper is technically sound 2
The paper describes original work 4
The work is of interest to the MedIA audience 4
The paper contains material which might well be omitted 5
The paper makes adequate references? 3
The abstract is an adequate digest of the work reported 3
The introduction gives the background of the work 2
The summary and conclusions adequate 2
The authors explain clearly what they have done 2
The authors explain clearly why what they did was worth doing 2
The order of presentation is satisfactory 3
The English is satisfactory 1
If there are color figures included, are they helpful/necessary? 3
If there is a video, is it helpful/necessary? N/A

Comments

The paper deals with the important topic of Diabetic Retinopathy (DR) early diagnosis. It presents the results of an international challenge hosted within the IEEE International Symposium on Biomedical Imaging in 2018. This challenge was organized in three subchallenges, each with a significant relevance: detection of DR lesions (microaneurysms, hemorrhages, hard exudates and cotton wool spots), location of retinal landmarks (fovea and optic disc, OD) and DR and diabetic macular edema (DME) severity grading. The methods proposed by the different teams were tested on the publicly available Indian
Diabetic Retinopathy Image Dataset (IDRiD). The best performing approaches for these three subchallenges and their results are presented in the paper.

The paper shows an interesting challenge in the context of DR diagnosis and grading. However, this Reviewer has some major issues regarding the manuscript.

1. Although the idea of the paper is relevant, the paper itself is very difficult to understand. First of all, authors need to thoroughly review the English and the style of the paper. I strongly recommend that authors have their paper reviewed by a native speaker.

2. Authors need to carefully review the style of the paper, especially if they want it to be published in a high impact journal like Medical Image Analysis. For example, they should use past simple whenever this is possible. Acronyms should be defined the first time they are used in the Abstract and manuscript text, and then they should always use the acronym. The style is not homogeneous throughout the text (this is especially notable in Appendix B, which I mention below). There is a general lack of references throughout the text, specially before equations. In some places, units that should accompany numbers are missing (for example, always use "pixels" or the adequate unit when referring to image sizes and "images" when referring to the number of images). Punctuation and the use of the article "the" should also be revised.

   Response: The paper has been carefully reviewed concerning these comments and revised thoroughly for the same.

3. In the Introduction section, authors need to better explain the challenge and the advantages of this challenge over previous existing ones. In my view, Table 1 should not be included here and would be better in Section 4.

   Response: We have incorporated this comment in the introduction section as per the following flow: Initially, mentioned the existing datasets → previous challenges in DR → cited the reviews that describe work done in the development of DR screening systems in the last two decades → Limitations of existing works → Finally, introduction of IDRiD dataset, the challenge and its advantage over existing ones.

➢ The following text (in blue) is included in the manuscript to address this comment:

“This necessity has led several research groups to develop and share retinal image datasets, namely Messidor (Decenciere et al., 2014), Kaggle (Cuadros and Bresnick, 2009), ROC (Niemeijer et al., 2010), E-Ophtha (Decenciere et al., 2013), DiaretDB (Kauppi et al., 2012), STARE (Hoover, 1975), ARIA (Farnell et al., 2008) and HEI-MED (Giancardo et al., 2012). Further, two challenges were organized in the context of DR, namely Retinopathy Online Challenge (ROC)2 and Kaggle DR detection challenge3. ROC was organized with the goal of detecting MAs. Whereas, the Kaggle challenge aimed to get solution for determining the severity level of DR. These challenges enabled advances in the field by promoting the participation of scientific research community from all over the globe on a competitive at the same time constructive setting for scientific advancement. Previous efforts have made good progress using image classification, pattern recognition, and machine learning. The progress through last two decades has been systematically reviewed by several research groups (Patton et al., 2006; Winder et al., 2009; Abramoff et al., 2010; Mookiah et al., 2013a; Jordan et al., 2017; Nørgaard and Grauslund, 2018).
Although lots of efforts have been made in the field towards automating the DR screening process, lesion detection is still a challenging task due to the following aspects: (a) Complex structures of the lesions (shape, size, intensity), (b) detection of lesions in tessellated images and in presence of noise (bright border reflections, impulsive noise, optical reflections), (c) high inter-class similarity (i.e. between MA-HE and EX-SE), (d) appearance of not so uncommon non-lesion structures (nerve fiber reflections, vessel reflections, drusens) and (e) difference in images obtained by different imaging devices makes it difficult to build a flexible and robust model for lesion segmentation. To the best of our knowledge, prior to the challenge, there were no reports on the development of a single framework to segment all lesions (MA, HE, SE, and EX) simultaneously. Also, there was a lack of common platform to test the robustness of approaches that determine the normal and abnormal retinal structures on the same set of images. Furthermore, there was limited availability of the pixel level annotations and the simultaneous gradings for DR and DME (see Tables in Appendix A).

In order to address these issues, we introduced a new dataset called Indian Diabetic Retinopathy Image Dataset (IDRiD) (Porwal et al., 2018a). Further, it was used as a base dataset for the organization of grand challenge on “Diabetic Retinopathy: Segmentation and Grading” in conjunction with ISBI - 2018. The IDRiD dataset provides expert markups of typical DR lesions and normal retinal structures. It also provides disease severity level of DR, and DME for each image in the database. This challenge brought together the computer vision and biomedical researchers with an ultimate aim to further stimulate and promote research, as well as to provide a unique platform for the development of a practical software tool that will support efficient and accurate measurement and analysis of retinal images that could be useful in DR management. Initially, a training dataset along with the ground truth was provided to participants for the development of their algorithms. Later, the results were judged on the performance of these algorithms on test dataset. Success was measured by how closely the algorithmic outcome matched the ground truth. There were three principal sub-challenges: lesion segmentation, disease severity grading, and localization and segmentation of retinal landmarks. These multiple tasks in IDRiD challenge allow to test the generalizability of the algorithms, and this is what makes it different from the existing ones. Further, this challenge seeks an automated solution to predict the severity of DR and DME simultaneously. It was projected as an individual task to increase the difficulty level of this challenge as compared to the Kaggle DR challenge i.e. for a given image, the predicted severity for both DR and DME should be correct to count for scoring the task.” (page no. 4 – 6 (line no 45 – 95))

➢ Further, we have moved Table 1 in section 4 (Now it appears as Table 5). This change could be observed at page no. 20.

4. Figure 1 is also in the reference Porwal et al. 2018b. Authors should make sure there is not a copyright problem.

Response: We have replaced Figure 1 with another image from the IDRiD dataset. This change could be observed at page no. 3.
5. In my opinion, the "Previous work" section should only mention the information that is relevant for the paper. In this sense, I think it could be combined with the "Introduction" section. In any case, authors should mention previous work related to the three sub-challenges (whether it is a deep learning-based approach or not), focusing on their advantages and disadvantages and how the proposed challenge can address some of the difficulties that arose in previous studies. Only the information relevant in this context should be mentioned in order to maintain focus.

Response: We have removed some theory detailing the retinal image analysis or ophthalmology (in general) and kept only the text specific to diabetic retinopathy. Even though we tried to compress the theory as much as it could be, however, considering the huge work done in the field (considering – three sub-challenges spread into eight subtasks of the challenge) it was not possible to mention previous work and combine it with the introduction section. Hence, we have mentioned the previous work related to the three sub-challenges in the separate section.

➢ The section related to previous work is titled “Review of Retinal Image Analysis for the detection of DR”. The following text (in blue) is included in the manuscript to address this comment. The underlined text represents the content that partly addresses this comment:

“Automatic image processing has proven to be a promising choice for the analysis of retinal fundus images and its application to future eye care. The introduction of automated techniques in DR screening programs and the interesting outcomes achieved by the rapidly growing deep learning technology are examples of success stories and potential future achievements. Particularly, after researcher’s (Krizhevsky et al., 2012) deep learning based model showed significant improvements over the state of the art in the ImageNet challenge, there was a surge of deep learning based models in medical image analysis. Hence, we decided to present the most recent relevant works with a classification based on whether or not they used deep learning in the context of DR.

2.1. Non-deep learning methods

The general framework for retinal image analysis through traditional handcrafted features based approaches involve several stages, typically: a preprocessing stage for contrast enhancement or non-uniformity equalization, image segmentation, feature extraction, and classification. The feature extraction strategy varies according to the objective involved i.e. retinal lesion detection, disease screening or landmark localization. In 2006, one research group (Patton et al., 2006) outlined the principles upon which retinal image analysis is based and discussed the initial techniques used to detect the retinal landmarks and lesions associated with DR. Later, one another group (Winder et al., 2009) reported an analysis of the work in the automated analysis of DR during 1998–2008. They categorized the literature into a series of operations or steps as preprocessing, vasculature segmentation, localization, and segmentation of the optic disk (OD), localization of the macula and fovea, detection and segmentation of lesions. Some of the review articles (Abramoff et al., 2010; Jordan et al., 2017) provide a brief introduction to quantitative methods for the analysis of fundus images with a focus on identification of retinal lesions and automated techniques for large scale screening for retinal diseases. Majority of attempts in the literature are towards exclusive detection and/or segmentation of one type of lesions (either MAs, HEs, EXs or SEs) from an image. Some of the common approaches involved for lesion
segmentation are mathematical morphology (Joshi and Karule, 2019; Hatanaka et al., 2008; Zhang et al., 2014), region growing (Fleming et al., 2006; Li and Chutatape, 2004), and supervised (Wu et al., 2017; Zhou et al., 2017; Garcia et al., 2009; Tang et al., 2013). Apart from these approaches, in case of MAs, most initial studies shown the effectiveness of template matching (Quellec et al., 2008), entropy thresholding (Das et al., 2015), radon space (Giancardo et al., 2011), sparse representation (Zhang et al., 2012; Javidi et al., 2017), hessian based region descriptors Adal et al. (2014), dictionary learning (Rocha et al., 2012). On the other hand, for exclusive segmentation of HEs, super-pixel based features (Tang et al., 2013; Romero-Oraa et al., 2019) were found to be effective. These red lesions (both MAs and HEs) are also frequently detected together using dynamic shape features (Seoud et al., 2016), filter response and multiple kernel learning (Srivastava et al., 2017) and hybrid feature extraction approach (Niemeijer et al., 2005). Similarly, for EXs researchers relied on approaches like clustering (Osareh et al., 2009), model-based (Sanchez et al., 2009; Harangi and Hajdu, 2014), ant colony optimization (ACO) (Pereira et al., 2015) and contextual information (Sanchez et al., 2012). Whereas, for SEs researchers utilized Scale Invariant Feature Transform (SIFT) (Naqvi et al., 2018), adaptive thresholding and ACO (Sreng et al., 2019). Further, several approaches were devised for multiple lesion detection such as multiscale amplitude-modulation-frequency-modulation (Agurto et al., 2010), machine learning (Roychowdhury et al., 2014), a combination of Hessian multiscale analysis, variational segmentation and texture features (Figueiredo et al., 2015). These techniques are shown to usually involve interdependence on the detection of anatomical structures (i.e. OD and fovea) with the lesion detection, and that in turn determines the automated DR screening outcome.

Localization and segmentation of OD and fovea facilitate the detection of retinal lesions as well as in the assessment (based on the geometric location of these lesions) of the severity and monitoring the progression of DR and DME. Hence, several approaches have been proposed for localization of OD, most of them utilized the OD properties like intensity, shape, color, texture, etc. and many others showed the effectiveness of mathematical morphology (Morales et al., 2013; Marin et al., 2015), template matching (Giachetti et al., 2014), deformable models (Yu et al., 2012; Wu et al., 2016) and intensity profile analysis (Kamble et al., 2017; Uribe-Valencia and Martinez-Carballido, 2019). Further, the approaches utilized for OD segmentation are based on level set (Yu et al., 2012), thresholding (Marin et al., 2015), active contour (Mary et al., 2015) and shape modeling (Cheng et al., 2015), clustering (Thakur and Juneja, 2017), and hybrid (Bai et al., 2014) approaches. Similarly, the fovea is detected mostly using the geometric relationship with OD and vessels through morphological (Welfer et al., 2011), thresholding (Gegundez-Arias et al., 2013), template (Kao et al., 2014) and intensity profile analysis (Kamble et al., 2017) techniques. Poor performance on the detection of normal anatomical structures could adversely affect lesion detection and screening accuracy. For instance, consider the mathematical morphology based techniques presented in 2002 (Walter et al., 2002), 2008 (Sopharak et al., 2008) and 2014 (Zhang et al., 2014). These works demonstrate how the morphological processing-based approaches evolved by including multiple steps for the final objective of exudate detection. In the initial efforts, Walter et al. devised a technique for OD and EXs segmentation, later removed the OD to obtain the exudate candidates. Similarly, Sopharak et al. achieved the same objective with the detection, and removal of OD and vessels. Recently, the approach presented by Zhang et al. achieved much better result, but it involved (a) spatial calibration, (b) detection of dark and bright
anatomical structures such as vessels and OD respectively, also (c) bright border regions detection before actual extraction of candidates. Also, there are other techniques based on textural (Morales et al., 2017; Porwal et al., 2018c) and mid-level (Pires et al., 2017) features of retinal images that forgo the lesion segmentation step for DR screening. However, most of these techniques depend on the intermediate steps mentioned above. In the approach based on machine learning (Roychowdhury et al., 2014) detected bright and dark lesions as a first step and later performed the hierarchical lesion classification to generate a severity grade for DR. Similarly, Antal and Hajdu (2014) proposed a strategy involving image-level quality assessment, pre-screening followed by lesion and anatomical features extraction to finally decide about the presence of DR using ensemble of classifiers. Further, for identification of different stages of DR features from morphological region properties (Yun et al., 2008), texture parameters (Acharya et al., 2012; Mookiah et al., 2013b), non-linear features of the higher-order spectra Acharya et al. (2008), hybrid Dhara et al. (2015) and information fusion (Niemeijer et al., 2009) approaches were found useful. As the DME is graded based on the location of the EXs from macula, many researchers (Giancardo et al., 2012; Medhi and Dandapat, 2014; Perdomo et al., 2016; Marin et al., 2018) proposed EXs based features to determine the severity of the DME. While several others (Deepak and Sivaswamy, 2012; Mookiah et al., 2015; Acharya et al., 2017) have proposed various feature extraction techniques to grade DME stages without segmenting EXs. Mainly for the approaches in this section, the features are based on the color, brightness, size, shape, edge strength, texture, and contextual information of pixel clusters in spatial and/or transform domain. Whereas the classification is achieved through the classifiers such as K Nearest Neighbors (KNN), Naive Bayes, Support Vector Machine (SVM), Artificial Neural Network (ANN), Decision Trees, etc.

These lesion detection or screening techniques are shown to usually involve interdependence with the other landmark detection. However, there is a lack of single platform to test their performance for each objective. For such handcrafted features based approaches this challenge provides a unique platform to compare and contrast the algorithm's performance for the detection of anatomical structures, lesions as well as screening of DR and DME.

2.2. Deep learning methods

Deep Learning is a general term to define multi-layered neural networks able to concurrently learn a low-level data representation and higher-level parameters directly from the data. This representation learning capability drastically reduces the need for engineering ad-hoc features, however, the full end-to-end training of deep learning based approaches typically require a significant number of samples. Its rapid development in recent times is mostly due to a massive influx of data, advances in computing power and developments in learning algorithms that enabled the construction of multilayer (more than two) networks (Hinton, 2018; Voulodimos et al., 2018). This progress has induced interests in the creation of analytical, data-driven models based on machine learning in health informatics (Ching et al., 2018; Raví et al., 2017). Hence, it is emerging as an effective tool for machine learning, promising to reshape the future of automated medical image analysis (Greenspan et al., 2016; Litjens et al., 2017; Suzuki, 2017; Shen et al., 2017; Kim et al., 2018; Ker et al., 2018). Among various methodological variants of deep learning, Convolutional Neural Networks (CNNs or ConvNets) are the most popular within the field of medical image analysis (Hoo-Chang et al., 2016; Carin and Pencina, 2018). Several configurations
and variants of CNN’s are available in the literature, some of the most popular are AlexNet (Krizhevsky et al., 2012), VGG (Simonyan and Zisserman, 2014), GoogLeNet (Szegedy et al., 2015) and ResNet (He et al., 2016).

Deep learning has also been widely utilized in the retinal image analysis because of its unique characteristic of preserving local image relations. Majority of the approaches in the literature employ deep learning to retinal images by utilizing “off-the-shelf CNN” features as complementary information channels to other handcrafted features or local saliency maps for detection of abnormalities associated with DR (Chudzik et al., 2018; Orlando et al., 2018; Dai et al., 2018), segmentation of OD (Zilly et al., 2017; Fu et al., 2018), and the detection of DR (Rangrej and Sivaswamy, 2017). The authors (Fu et al., 2016) employ fully connected conditional random fields along with CNN to integrate the discriminative vessel probability map and long-range interactions between pixels to obtain final binary vasculature. Whereas some approaches initialized the parameters with those of pre-trained models (on non-medical images), then “fine-tuned” (Tajbakhsh et al., 2016) the network parameters for DR screening (Gulshan et al., 2016; Carson Lam et al., 2018). In another approach researchers used two-dimensional (2D) image patches as an input instead the full-sized images for lesion detection (Tan et al., 2017b; van Grinsven et al., 2016; Lam et al., 2018; Chudzik et al., 2018; Khojasteh et al., 2018), and OD and fovea detection (Tan et al., 2017a). In (Garcia et al., 2017) trained the “CNN from scratch” and compared it with the finetuning results based on the other two existing architectures. Recently, Shah et al. (2018) demonstrated that the ensemble training of auto-encoders stimulates diversity in learning dictionary of visual kernels for detection of abnormalities. Whereas Giancardo et al. (2017) proposed a novel way to compute the vasculature embedding that leverages the internal representation of a new encoder-enhanced CNN, demonstrating improvement in the DR classification and retrieval task.

There is a significant development in the automated identification of DR using CNN models in recent time. A customized CNN (Gargeya and Leng, 2017) proposed for DR screening and trained using 75,137 obtained from EyePACS system (Cuadros and Bresnick, 2009), where an additional classifier was further employed on the CNN-derived features to determine if the image is with or without retinopathy. Similarly, Google Inc. (Gulshan et al., 2016) developed a network optimized (fine tuning) for image classification, in which a CNN is trained by utilizing a retrospective development database consisting of 128,175 images with the labels. There are some hybrid algorithms, in which multiple, semi-dependent CNN’s are trained based on the appearance of retinal lesions (Abramoff et al., 2016; Quellec et al., 2016). A step further, the researchers (Quellec et al., 2017) demonstrated an ability of lesion segmentation based on the CNN trained for image level classification. However, Lynch et al. (2017) demonstrated that the hybrid algorithms based on multiple semi-dependent CNNs might offer a more robust option for DR referral screening, stressing the importance of lesion segmentation. For further details, readers are recommended to follow recent reviews for detection of exudates (Fraz et al., 2018), red lesions (Biyani and Patre, 2018) and a systematic review with a focus on the computer-aided diagnosis of DR (Mookiah et al., 2013a; Nørgaard and Grauslund, 2018).

This current progress in artificial intelligence provides an opportunity to the researchers for enhancing the performance of the DR referral system to more robust diagnosis system that can provide the quantitative information for multiple diseases matching the international
standards of clinical relevance. Thus, this challenging design offers an avenue to gauge precise DR severity status and opportunity to deliver accurate measures for lesions, that could even help in the follow-up studies to observe changes in the retinal atlas.” (page no. 7 – 12 (line no 106 – 282))

➢ This comment has also been partly taken care while incorporating with the comment no 3 by mentioning the limitation of existing works and state how the proposed challenge can address some of the difficulties as follows:

“Although lots of efforts have been made in the field towards automating the DR screening process, lesion detection is still a challenging task due to the following aspects: (a) Complex structures of the lesions (shape, size, intensity), (b) detection of lesions in tessellated images and in presence of noise (bright border reflections, impulsive noise, optical reflections), (c) high inter-class similarity (i.e. between MA-HE and EX-SE), (d) appearance of not so uncommon non-lesion structures (nerve fiber reflections, vessel reflections, drusen) and (e) difference in images obtained by different imaging devices makes it difficult to build a flexible and robust model for lesion segmentation. To the best of our knowledge, prior to the challenge, there were no reports on the development of a single framework to segment all lesions (MA, HE, SE, and EX) simultaneously. Also, there was a lack of common platform to test the robustness of approaches that determine the normal and abnormal retinal structures on the same set of images. Furthermore, there was limited availability of the pixel level annotations and the simultaneous gradings for DR and DME (see Tables in Appendix A).

In order to address these issues, we introduced a new dataset called Indian Diabetic Retinopathy Image Dataset (IDRiD) (Porwal et al., 2018a). Further, it was used as a base dataset for the organization of the grand challenge on “Diabetic Retinopathy: Segmentation and Grading” in conjunction with ISBI - 2018. The IDRiD dataset provides expert markups of typical DR lesions and normal retinal structures. It also provides disease severity level of DR, and DME for each image in the database. This challenge brought together the computer vision and biomedical researchers with an ultimate aim to further stimulate and promote research, as well as to provide a unique platform for the development of a practical software tool that will support efficient and accurate measurement and analysis of retinal images that could be useful in DR management. Initially, a training dataset along with the ground truth was provided to participants for the development of their algorithms. Later, the results were judged on the performance of these algorithms on the test dataset. Success was measured by how closely the algorithmic outcome matched the ground truth. There were three principal sub-challenges: lesion segmentation, disease severity grading, and localization and segmentation of retinal landmarks. These multiple tasks in IDRiD challenge allow to test the generalizability of the algorithms, and this is what makes it different from the existing ones. Further, this challenge seeks an automated solution to predict the severity of DR and DME simultaneously. It was projected as an individual task to increase the difficulty level of this challenge as compared to the Kaggle DR challenge i.e. for a given image, the predicted severity for both DR and DME should be correct to count for scoring the task.” (Page no. 5-6 (line no: 60 – 95))
6. Although IDRiD database has some advantages over previously published public databases, there is no need to deeply describe those databases in order to highlight the benefits of IDRiD. In my opinion, Appendix A should be removed and only the relevant references included in the paper, embedded in the manuscript text. In this sense, Tables 2 and 3 could be much simplified, and maybe authors could refer readers to Porwal et al. 2018a for some of the details. Please note that the aim of this study was to use only the IDRiD database.

Response: We have removed Appendix A detailing previous datasets and relevant references are included in the introduction section of this paper. Also, we moved Table 2 and 3 to the appendix (Now appear as Table A1 and A2 as shown below) so the interested readers could refer them for more details. This change could be observed on page no. 59.

Table A.1. Summary of technical specifications and hardware used in different databases

<table>
<thead>
<tr>
<th>Name of Database</th>
<th>Number of Images</th>
<th>Image Size(s)</th>
<th>FOV</th>
<th>Camera</th>
<th>NMY</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARIA</td>
<td>212</td>
<td>768×576</td>
<td>50</td>
<td>Zeiss</td>
<td>✓</td>
<td>TIFF</td>
</tr>
<tr>
<td>DIARETDB</td>
<td>130+89</td>
<td>1500×1152</td>
<td>50</td>
<td>Zeiss</td>
<td>✓</td>
<td>PNG</td>
</tr>
<tr>
<td>DRIVE</td>
<td>40</td>
<td>768×584</td>
<td>45</td>
<td>Canon CR5</td>
<td>✓</td>
<td>JPEG</td>
</tr>
<tr>
<td>E-Ophtha</td>
<td>47EX+35H, 148MA+233H</td>
<td>1440×960 - 2048×1360 (4)</td>
<td>45</td>
<td>Canon CR - DG1 &amp; Topcon TRC - NW6</td>
<td>✓</td>
<td>JPEG</td>
</tr>
<tr>
<td>HEIMED</td>
<td>169</td>
<td>2196×1958</td>
<td>45</td>
<td>Zeiss Visucam PRO</td>
<td>✓</td>
<td>JPEG</td>
</tr>
<tr>
<td>Kaggle</td>
<td>88,702</td>
<td>Varying</td>
<td></td>
<td>Any camera (EyePACS Platform)</td>
<td>-</td>
<td>TIFF</td>
</tr>
<tr>
<td>MESSIDOR</td>
<td>800 MY+ 400 NMY+ 1756</td>
<td>1440×960, 2240×1488, 2304×1536</td>
<td>45</td>
<td>3CCD/ Topcon TRC NW6</td>
<td>Both</td>
<td>TIFF</td>
</tr>
<tr>
<td>ROC</td>
<td>100</td>
<td>768×576, 1058×1061, 1389×1383</td>
<td>45</td>
<td>Topcon NW 100 &amp; NW 200, Canon CR5 - 45NM</td>
<td>✓</td>
<td>JPEG</td>
</tr>
<tr>
<td>STARE</td>
<td>397</td>
<td>605×700</td>
<td>35</td>
<td>Topcon TRV - 50</td>
<td>×</td>
<td>PPM</td>
</tr>
<tr>
<td>IDRiD</td>
<td>516 (81 with LA)</td>
<td>4288×2848</td>
<td>50</td>
<td>Kowa VX - 10α</td>
<td>✓</td>
<td>JPG</td>
</tr>
</tbody>
</table>

7. When describing the IDRiD database, please include the image capture protocol (how many images per eye were captured, where were they centered...). Please briefly describe the "International Clinical Diabetic Retinopathy Scale" used for DR and DME grading (and provide a relevant reference). It is also unclear how the OD boundary was delineated (authors only mention that OD and fovea centers were marked, but a subtask regarding the complete OD segmentation is also included in the challenge). Please explain how the image set was divided into training and test subsets (randomly?) and how the percentages of the images that should be in each subset were chosen. Does the database include images without any lesion?

**Response:** As per the recommendations, we have included the details regarding image capture protocol, severity grading for DR and DME (Table 1 and 2 on page no. 14), and division of training and test set (Table no. 3 and 4 on page no. 16-17). Further, the information about OD delineation is presented in subsection – ‘Pixel level annotations’ and the same is illustrated in Figure 2(f) (page no. 13). Further, the explanation regarding the data division and percentages is included on page no. 15-17 (line no. 338-351). The database includes 168 images without lesion as shown in Table 4 under "Grade – 0" (set A + set B).

➢ The following text (in blue) is included in the manuscript to address this comment:

"The fundus photographs of people affected by diabetes were captured with focus on macula using Kowa V X-10α fundus camera. Prior to capturing of images, pupils of all subjects were dilated with one drop of tropicamide at 0.5% concentration. The captured images have 50° field of view and resolution of 4288 x 2848 pixels stored in jpg format.” (page no. 12-13 (line no. 287 – 291))
“The diabetic retinal images were classified into separate groups according to the International Clinical Diabetic Retinopathy Scale (Wu et al., 2013) as shown in Table 1. The DME severity was decided based on occurrences of EXs near to macula center region (Decenciere et al., 2014) as shown in Table 2.” (page no. 14 (line no. 307 – 311))

“1. Pixel Level Annotations. This type of annotations are useful in the techniques to locate individual lesions within an image and to segment out regions of interest from the background. Eighty-one color fundus photographs with signs of DR are annotated at pixel level for developing ground truth of MAs, SEs, EXs and HEs. The binary masks (as shown in Fig. 2) for each type of lesion are provided in tif file format. Additionally, OD was also annotated at pixel level and binary masks for all 81 images are provided in the same format. “(page no. 13 (line no. 299-303))

“The dataset along with the groundtruths were separated into training set and test set. For the images with pixel level annotations the data was separated as 2/3 for training (Set-A) and 1/3 for testing (Set-B) (See Table 3). Similarly, data for the OD segmentation (part of sub-challenge – 3) was divided in same ratio into Set-A (54 images) and Set-B (27 images). The percentage of images that should be in each subset for lesion and OD segmentation tasks (sub-challenge – 1 and part of sub-challenge – 3) were chosen based on the research outcome (Dobbin and Simon, 2011) which demonstrated that splitting data into 2/3 (training): 1/3 (testing) is an optimal choice for the sample sizes from 50 to 200. For the other sub-challenges (disease grading, and OD and fovea center locations), data was separated in 80 (training set: Set-A): 20 (testing set: Set-B) ratio. The percentage of data split in this case is done to provide an adequate amount of data divided into different severity levels. Note that the dataset was stratified according the DR and DME grades before splitting. A breakdown of the details of the dataset is shown in Table 4.” (page no. 15-17 (line no. 338 – 351))

8. In section 4, "Challenge organization", please make sure that the different stages described match Figure 3.  
Response: We have modified the challenge organization section and made sure that the different stages described match Figure 3 (page no. 15-19).

9. I still have some doubts regarding the challenge organization. First of all, authors claim that participants could submit "up to three methods"; but I don’t know if that means that they could only, for example, detect three of the four types of lesions in subchallenge 1 (lesion segmentation). I believe that is not the case because team iFLYTEK detect the four lesion types, but please clarify this issue. Regarding subchallenges 1 and 3, teams could decide to participate only in some of the Tasks, is that correct? Why was that not allowed in subchallenge 2 (i.e. to detect only DR or DME severity)?

Response: We have incorporated this comment in the manuscript while detailing the challenge organization. Here we initially introduced three challenges divided into eight tasks and then mentioned that participants could submit up to three methods to be evaluated per team for each task.
The following text (in blue) is included in the manuscript to address this comment:

“Participants could submit up to three methods to be evaluated per team for each task, provided that there was a significant difference between the techniques, beyond a simple change or alteration of parameters.” (page no. 15-17 (line no. 387 – 389))

In case of sub-challenge-2, the choice to detect DR and DME simultaneously is explained in section 1 on page no. 6 (line no. 92-96) and its reason is explained in section 6 while detailing the performance evaluation measures for subchallenge-2 on page no. 43 (line no. 1007-1013).

“Further, this challenge seeks an automated solution to predict the severity of DR and DME simultaneously. It was projected as an individual task to increase the difficulty level of this challenge as compared to the Kaggle DR challenge i.e. for a given image, the predicted severity for both DR and DME should be correct to count for scoring the task.” (page no. 6 (line no. 92-96))

“This was done since, even with presence of some exudation that may be categorized as mild DR, its location on the retina is also important governing factor (to check DME) to decide overall grade of disease. For instance, EXs presence in the macular region can affect vision of the patient to greater extent and hence, it should be dealt with priority for referral (that may otherwise be missed or cause delay in treatment with the present convention of only DR grading) in the automated screening systems.” (page no. 43 (line no. 1016-1022))

10. In sub-challenge 1, since the evaluation on the test set was done off-line. How did organizers ensure that results were measured in the same way by all teams?

Response: Participants were asked to submit all output images/csv files along with the short paper describing the technical details and then all results were evaluated by the organizers. We have addressed this comment in section 4.

The following text (in blue) is included in the manuscript to address this comment:

“For Tasks 1 to 4 (i.e. subchallenge – 1) and task-8, the teams were asked to submit output probability maps as grayscale images and for all other tasks it was accepted in CSV format. The submitted results were evaluated by the challenge organizers and their performance was displayed on leaderboard of the challenge website.” (page no. 18 (line no 389-393))

11. In sub-challenge 3, task 8. What was the ground truth for teams?

Response: To address this comment we have included text detailing it in the section 4. Binary images (as shown in Fig. 2(f)) in tif format was ground truth for the teams.

The following text (in blue) is included in the manuscript to address this comment:

“Additionally, OD was also annotated at pixel level and binary masks for all 81 images are provided in the same format.” (page no. 16 (line no. 301-303))

“For the images with pixel level annotations the data was separated as two third for training (Set-A) and one third for testing (Set-B) (See Table 3). Similarly, data for the OD segmentation (part of sub-challenge – 3) was divided in same ratio into Set-A (54 images) and Set-B (27 images).” (page no. 15-16 (line no. 339-341))
12. Authors need to better organize the information regarding the participating methods and to better explain the different approaches. In the text, authors should give only the relevant details regarding the methods proposed by the different teams. However, the explanations need to be sufficient for a non-expert reader to follow the ideas of the paper (for example, all the relevant terminology should be described). For readers who would like a more comprehensive description of one particular method, relevant references should be provided. This way, the paper is understandable and, at the same time, the focus on the topic of the paper is maintained. However, the explanations need to be sufficient for a non-expert reader to follow the ideas of the paper (for example, all the relevant terminology should be described).

13. In this sense, I believe Appendix B is not adequate in this paper. In Appendix B the methods are not thoroughly described (it would be implausible), so readers do not really get a comprehensive view of the methods and there are a lot of terms and concepts that are not understandable for a reader not familiarized with the method. Thus, it would be much better if readers could refer to a relevant reference if they are interested in a particular method. Besides, the description of the methods in Appendix B is quite variable. It appears as if each team had written something on their method separately and that was just copy-pasted in Appendix B, without giving it any kind of uniformity (references, acronyms, ...). Thus, in my view, both Appendix A and Appendix B should be removed and only the relevant information on these Appendices included within the manuscript text. Please note that this makes 24 pages of the manuscript that, in my opinion, distract the attention of readers from the relevant topic of the manuscript: the DR diagnosis challenge.

Response (for comments 12 and 13): We have initially described the required theory to give a comprehensive view of methods and built upon that theory to summarize the participating solutions. We have removed both Appendix A and Appendix B. Further, for the readers who are interested to know the complete details of a particular solution, the link to full papers of all participating teams is provided on the challenge website at https://idrid.grand-challenge.org/Challenge_Proceedings/.

➢ The following text (in blue) is added in the manuscript to describe the theory required to give a comprehensive view of methods:

“Majority of participating teams proposed a CNN based approach for solving tasks in this challenge. This section details the basic terminologies and abbreviations related to CNN and its variants utilized by the participating teams. Further it summaries the solutions and related technical specifications. For the detailed description of a particular approach please refer to the proceedings of the ISBI Grand Challenge Workshop at https://idrid.grand-challenge.org/Challenge_Proceedings/.

For the input image, CNN transforms the raw image pixels on one end to generate a single differentiable score function at the other. It exploits three mechanisms — sparse connections (a.k.a. local receptive field), weight sharing and invariant (or equivariant) representation — that makes it computationally efficient (Shen et al., 2017). The CNN architecture typically consists of an input layer followed by sequence of convolutional (CONV), subsampling (POOL), fully-connected (FC) layers and finally a SoftMax or regression layer, to generate the desired output. Functions of all layers are detailed as follows:
The CONV layer comprises of a set of independent filters (or kernels) that are utilized to perform 2D convolution with the input layer \((I)\) to produce the feature (or activation) maps \((A)\) that give the responses of kernels at every spatial position. Mathematically, for the input patch \((I_{x,y}^\ell)\) centered at location \((x,y)\) of the \(\ell^{th}\) layer, the feature value in the \(i^{th}\) feature map, \(A_{x,y,i}^\ell\), is obtained as:

\[
A_{x,y,i}^\ell = f((w_i^\ell)_{x,y} I_{x,y}^\ell + b_i^\ell) = f(C_{x,y,i}^\ell)
\]

Where the parameters \(w_i^\ell\) and \(b_i^\ell\) are weight vector and bias term of the \(i^{th}\) filter of the \(\ell^{th}\) layer, and \(f(\cdot)\) is a nonlinear activation function such as sigmoid, rectified linear unit (ReLU) or hyperbolic tangent (tanh). It is important to note that the kernel \(w_i^\ell\) that generates the feature map \(C_{x,y,i}^\ell\) is shared, reducing the model complexity and making the network easier to train.

The POOL layer aims to achieve translation-invariance by reducing the resolution of the feature maps. Each unit in a feature map of the POOL layer is derived using a subset of units within sparse connections from the corresponding convolutional feature map. The most common pooling operations are average pooling and max pooling. It performs downsampling operation and is usually placed between two CONV layers to achieve a hierarchical set of image features. The kernels in the initial CONV layers detect low-level features such as edges and curves, while the kernels in the higher layers are learned to encode more abstract features. The sequence of several CONV and POOL layers gradually extract higher-level feature representation.

FC layer aims to perform higher-level reasoning by computing the class scores. Each neuron in this layer is connected to all neurons in the previous layer to generate global semantic information.

The last layer of CNN’s is an output layer \((O)\), here the Soft-Max operator is commonly used for the classification tasks. The optimum parameters \((\Theta)\), common notation for both \(w\) and \(b\) for a particular task can be determined by minimizing the loss function \((L)\) defined for the task. Mathematically, for \(N\) input-output relations \({(I^n, O^n); \ n \in [1, \ldots, N]}\) and corresponding labels \(G^n\) the loss can be derived as:

\[
L = \frac{1}{N} \sum_{n=1}^{N} \ln(\Theta; G^n, O^n))
\]

Where \(N\) denotes the number of training images, \(I^n, O^n\) and \(G^n\) correspond to the \(n^{th}\) training image. Here, a critical challenge in training CNN’s arises from the limited number of training samples as compared to the number of learnable parameters that need to be optimized for the task at hand. Recent studies have developed some key techniques to better train and optimize the deep models such as data augmentation, weight initialization, Stochastic Gradient Descent (SGD), batch normalization, shortcut connections and regularization. For more understanding related to advances in CNN’s, reader is recommended to refer (Gu et al., 2018).

The growing use of CNN’s as the backbone of many visual tasks, ready for different purposes (such as segmentation, classification or localization) and available data, has made architecture search a primary channel in solving the problem.

In this challenge, mainly for disease severity grading problem, participants either directly utilized existing variants of CNN’s or ensembled them to demarcate the input image to one of the class
mentioned above. Several configurations and variants of CNN's are available in literature, some of the most popular are AlexNet (Krizhevsky et al., 2012), VGG (Simonyan and Zisserman, 2014), GoogLeNet (Szegedy et al., 2015) and ResNet (He et al., 2016) due to their superior performance on different benchmarks for object recognition tasks. A typical trend with the evolution of these architectures is that the networks have gotten deeper, e.g., ResNet is about 19, 8 and 7 times deeper than AlexNet, VGGNet, and GoogLeNet respectively. While the increasing depth improves feature representation and prediction performance, it also increases complexity, making it difficult to optimize and even becomes prone to overfitting. Further, the increasing number of layers (i.e., network depth) leads to vanishing gradient problems as a result of a large number of multiplication operations. Hence, many teams chose the DenseNet (Iandola et al., 2014) which connects each layer to every other layer in a feed-forward fashion, reducing the number of training parameters and alleviates the vanishing gradient problem. DenseNet exhibits $\ell (\ell + 1)/2$ connections in $\ell$ layer network, instead of only $\ell$, as in the networks mentioned above. This enables feature reuse throughout the network that leads to more compact internal representations and in turn, enhances its prediction accuracy. Another opted approach, Deep Layer Aggregation (DLA) structures (Yu et al., 2017), extends the “shallow” skip connections in DenseNet to incorporate more depth and sharing of the features. DLA uses two structures – iterative deep aggregation (IDA) and hierarchical deep aggregation (HDA) that iteratively and hierarchically fuse the feature hierarchies (i.e. semantic and spatial) to make networks work with better accuracy and fewer parameters. Recent Fully Convolutional Network (FCN) (Long et al., 2015) adapt and extend deep classification architectures (VGG and GoogLeNet) into fully convolutional networks and transfer their learned representations by fine-tuning to the segmentation task. It defines a skip architecture that combines semantic information from a deep, coarse layer with appearance information from a shallow, fine layer to produce accurate and detailed segmentations.

For the lesion segmentation task, most of the participating teams exploit U-Net architecture (Ronneberger et al., 2015). The main idea in U-Net architecture is to supplement the usual contracting network through a symmetric expansive path by addition of successive layers, where upsampling (via deconvolution) is performed instead of pooling operation. The upsampling part consists of large number of feature channels, that allow the network to propagate context information to higher resolution layers. The high-resolution features from the contracting path are merged with the upsampled output and fed to soft-max classifier for pixel-wise classification. This network works with very few training images and enables the seamless segmentation of high-resolution images by means of an overlap-tile strategy. Other similar architecture SegNet (Badrinarayanan et al., 2015) was opted by a team, it consists of an encoder and decoder network, where the encoder network is topologically identical to the CONV layers in VGG16 and in which FC layer is replaced by a SoftMax layer. Whereas, the decoder network comprises a hierarchy of decoders, one corresponding to each encoder. The decoder uses max-pooling indices for upsampling its encoder input to produce a sparse feature map. Later, it convolves the sparse feature maps with a trainable filter bank to densify them. At last, the decoder output is fed to a soft-max classifier for generation of segmentation map. One team choose Mask R-CNN (He et al., 2017), a technique primarily based on a Region Proposal Network (RPN) that shares convolutional features of entire image with the detection network, thus enabling region proposals to localize and further segments normal and abnormal structures in the retina. RPN is a fully convolutional
network that contributes in concurrently predicting object bounds and “objectness” scores at each position.

Following subsections present the solutions designed by participating teams with respect to three sub-challenges. Table 6 summarizes the data augmentation, normalization and preprocessing tasks performed by each team.” (page no. 15-24 (line no. 426-533))

➢ After this text all approaches that were in the appendix B are revised and included in this section (page no. 25-42) (it is divided into three subsections respectively for the three sub-challenges).

14. Regarding evaluation measures, please justify better the choices for each subchallenge. Please provide references for the different measures used. I also find that authors need to explain what each of the participating teams had to send for evaluation in each subchallenge (binary images? Images with a probability map? Csv files? Other?).

Response: We have detailed the choices of performance measures used for each sub-challenge and provided references for the same. We have explained result formats in the section 4 (page no.18, line no. 389-391) and also mentioned them in Section 6 while detailing performance measures. This change could be observed on page nos. 42 and 44.

➢ The following text (in blue) is included in the manuscript to address these comments:

“A. Sub-challenge – 1

This sub-challenge evaluates the performance of the algorithms for different lesion segmentation tasks, from the submitted grayscale images, using the available binary masks. As in the lesion segmentation task(s) background overwhelms foreground, a highly imbalanced scenario, the performance of this task was measured using area under precision (a.k.a. Positive Predictive Value (PPV)) recall (a.k.a. Sensitivity (SN)) curve (AUPR) (Saito and Rehmsmeier, 2015). The AUPR provides a single-figure measure (a.k.a. mean average precision (mAP)), computed over the set-B, was used to rank the participating methods. This performance metric was used for object detection in The PASCAL Visual Object Classes (VOC) Challenge (Everingham et al., 2010). The AUPR measure is more realistic (Boyd et al., 2013; Saito and Rehmsmeier, 2015) for the lesion segmentation performance over the Area under Receiver Operating Characteristics.”

“B. Sub-challenge – 2

Let the expert labels for DR and DME be represented by DR_G(n) and DME_G(n). Whereas, DR_O(n) and DME_O(n) are the predicted results, then correct instance is the case when the expert label for DR and DME matches with the predicted outcomes for both DR and DME. This was done since, even with presence of some exudation that may be categorized as mild DR, its location on the retina is also important governing factor (to check DME) to decide overall grade of disease. For instance, EXs presence in the macular region can affect vision of the patient to greater extent and hence, it should be dealt with priority for referral (that may otherwise be missed or cause delay in treatment with the present convention of only DR grading) in the automated screening systems. Hence, disease grading performance accuracy for this sub-
challenge, from the results submitted in CSV format for test images (i.e. $N = 103$), is obtained by algorithm 1 as follows:"

"C. Sub-challenge – 3

For the given retinal image, the objective of sub-challenge-3 (task – 6 and 7) was to predict the OD and fovea center co-ordinates. The performance of results submitted in CSV format was evaluated by estimating the Euclidean distance (ED) (in pixels) between manual (ground truth) and automatically predicted center location. Lower ED indicates better localization. After determining Euclidean distance for each image in the set-B, i.e. for 103 images, the average distance representing the whole dataset was computed and used to rank the participating methods. The optic disc segmentation (task – 8) performance is evaluated using Jaccard index ($J$) (Jaccard, 1908). It represents the proportion of overlapping area between the segmented OD ($A$) and the ground truth ($B$). Higher $J$ indicates better segmentation. For the segmented results, images in range $[0, 255]$, it was computed at 10 different equally spaced thresholds $[0, 0.1, \cdots, 0.9]$ and averaged to obtain final score."

15. In the results of Subchallenge 1 authors claim that the teams were ranked according to their performance on each type of lesion and to their "overall performance". How was the latter measured?

Response: All four tasks in Subchallenge-1 are considered individually and hence ranked independently. The term “overall performance” could be clear from the Table given below. Here it means the solutions developed by the teams that ranked amongst the top three for at least three different lesion segmentation tasks, presented their work in the ISBI workshop. We have addressed this comment in the paper for more clarity.

![Table showing team names, scores, and ranks for each task.]

➢ The following text (in blue) is included in the manuscript to address these comments:

“Amongst them, only top-4 teams per lesion segmentation task were invited for the challenge workshop and top-3 teams having overall better performance, the solutions developed by the teams that ranked amongst top three for at least three different lesion segmentation tasks, presented their work at ISBI.” (page no. 44 (line no. 1049-1053))
16. In Figure 4 I would recommend authors to include the results of the 4 top-teams on each type of lesion (not only 3...).

**Response:** We have included the results of 4 top-teams on each type of lesion (Now Figure 6). To maintain uniformity, we have also included results of the 4 top-teams for OD segmentation (Figure 9). This change could be observed on page no. 46 and 50.

17. In figure 5, please include figure legends in a bigger font size. Please include the different approaches in the same order in the different sub-figures. In my view, authors do not need to include AUC in the legend, since this information is already in Table 7.

**Response:** We have modified Figure 5 (Now Figure 7) to appear clearer, removed AUC in the legend, and included different approaches in the same order. This change could be observed on page no. 47.

18. Please discuss further the results of subchallenge 3. In my view, the performance of the methods is very related to the image resolution employed by each team (i.e., the same Euclidean distance, in pixels, between the detected OD or fovea center and the ground truth does not mean the same in a "bigger" image than in a "smaller" image). Please discuss this issue. I would find it very useful to include the average OD diameter for each team (or image resolution) since it may give readers a better understanding of the performance of the different methods. Indeed, in many studies, the detection of the OD or fovea center is considered correct if it is less than an OD radius apart from the center annotated by experts.

**Response:** We have discussed the results of all sub-challenges in relation to the image resolution employed by each team. The content added in response to this comment also partly incorporates the comment no. 20. The following text (in blue) is added in the manuscript to address this comment:

“As expected, we found that image resolution is a vital factor for the model performance, especially for the task of segmentation of small objects such as MAs or EXs. In fact, the top performing approaches process the images patch-wise, which allow models to have a local high-resolution image view or directly with the high-resolution image as a whole. This is essential as MAs or small EXs lesions span very few pixels in some cases and reducing the original image size would prevent an accurate segmentation. Similarly, image resolution plays a very important role for the disease classification task (see Table 9), the most likely reason is that the presence of the disease is determined by the presence of lesions in the image, including the small ones that might be invisible at low resolution. This is corroborated by the confusion matrices in Table 12 which show misclassified instances in DR (particularly, grade 1 and 2) as well as DME (5 images each belonging to grade 1 and 2 are predicted as grade 0). For the localization tasks, all participants were asked to identify retinal structures with coordinates at full image resolution. Most of them performed these tasks by scaling the image to smaller size and then converted their predictions in the original image space. The results indicate that the input image resolution has limited effect on the results of the localization problem. For instance, in case of OD localization, the top performing team utilized two image resolutions, one (224 × 224 pixels) for approximate location prediction and other (cropped ROIs 950 × 950 pixels) for refining that estimate. Similarly, teams CBER and VRT resized the image to 536 × 356 pixels and 640 × 640 pixels respectively to get an
approximate center location whereas, the team SDNU utilized the input size of 1984 × 1318 pixels. Considering the OD average diameter of approximately 516 pixels, the deflection of result for about 10 to 15 pixels by other approaches, utilizing different input resolutions, as compared to the top performing solution is very less. This is because the retinal structures to be identified, OD and fovea, are very unlikely to disappear due to a reduction of image resolution and they have clear geometrical constraints.” (page no. 54 and 56 (line no. 1165-1191))

➢ Further, we have computed the average OD diameter of all images in the test set for all competing teams. A figure illustrating the performance of each team with respect to ground truth is presented in the discussion section.

The following text in blue is added in the manuscript to address this comment:

“Considering the clinical significance of OD diameter while DME severity grading, we further compute the average OD diameter (in pixels) for each image of test set. Figure 13 illustrates the performance of each participating team with respect to the groundtruth, most methods show a similar pattern. The average diameter of OD groundtruth is 516.61 pixels whereas, this corresponding values for the results of solution developed by the teams ZJU-BII-SGEX, VRT, IITKgpKLIV, CBER and SDNU are 514.25, 519.21, 513.48, 508.04 and 460.19 pixels respectively.” (page no. 54 - 55 (line no. 1156-1162))

![Figure 13: Illustration of average OD diameter result of all 5 teams for each image of the testing dataset. [Here the legends G, ZJU and IITKgp represent Groundtruth, ZJU-BII-SGEX and IITKgpKLIV respectively (compressed to appear clearly in single column format, appears full in double column format.).] ](image)

19. Authors do not need to explain what a boxplot is (and definitely not in a footnote). Please substitute that for a relevant reference.

**Response:** We have substituted the explanation of boxplot with a relevant reference. This change could be observed on page no. 49 (line no 1094).
20. I would recommend authors to extend their discussion. It may be relevant to explain (for the 3 subchallenges) the cases where the proposed methods tended to fail or those where they normally performed well. I would also recommend authors to discuss the clinical relevance of this challenge.

**Response:** We have extended the discussion to present the successful and failure cases. We have also presented the clinical relevance of this challenge in the introduction section of thoroughly revised manuscript. This change could be observed on page no. 52 (inclusion of Figures 11 and 12 highlighting successful and failure cases respectively), 53 (inclusion of confusion matrices (Table 12)).

➢ The following text (in blue) is added in the manuscript to address this comment:

“Fig. 11 highlights the performance of top solution for EX that performs significantly well in presence of normal retinal structures and different challenging circumstances.” (line no. 1113-1115)

![Figures 11 and 12](image_url)

Fig. 11. Illustration of (a-d) different challenging circumstances for segmentation of EXs, (e-f) segmentation results (probability map) of top-performing team for EXs, (i) enlarged part of Fig. (d), and (j) depicts its performance to be better than (k) the human annotator (The annotator tool had limitation of the markup capability when there is an overlap of multiple types of lesion. In this case, EXs and HE).

“Further, Fig. 12 shows that some false positives detected by the participating solutions are due to noise, predominantly for MA and HE. This indicates that there is still room for improvement for lesion segmentation tasks with current fundus cameras.” (line no. 1113-1115)
Considering the misclassified instances in the confusion matrices in Table 12, along with the lesion information, it is essential to give attention towards characterization of intra-retinal microvascular abnormalities (IRMA’s) and venous beading for improvement in the overall grading results."

➢ Apart from this content we have also discussed the reasoning behind the success and failure of solutions with respect to the input image resolution as presented in response to comment no. 18.

➢ Further the tasks included in this challenge are supported by relevant explanations in introduction, literature review and performance evaluation sections with the text as follows:

“Programs to screen such a large population for DR confront the issues related to the implementation, management, availability of human graders, and long-term financial sustainability. Hence, computer aided diagnosis tools are required for screening such a large population that require continuous follow-up for DR and to effectively facilitate in reducing the burden on the ophthalmologists (Jelinek and Cree, 2009; Walter et al., 2002). Such a tool would help clinicians in the identification, interpretation, and measurements of retinal abnormalities, and ultimately in the screening and monitoring of the disease.”

Fig. 12: Illustration of results by top performing solutions for (a-c) different images with noise causing most common false positives in the segmentation of (d-f) MAs, and (g-i) HEs respectively.

Table 12: Confusion matrix of retinal images predicted by top performing solution for DR (5 class) and DME (3 class).

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>30</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>22</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>40</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>2</td>
<td>41</td>
<td></td>
</tr>
</tbody>
</table>

➢ Apart from this content we have also discussed the reasoning behind the success and failure of solutions with respect to the input image resolution as presented in response to comment no. 18.

➢ Further the tasks included in this challenge are supported by relevant explanations in introduction, literature review and performance evaluation sections with the text as follows:

“Programs to screen such a large population for DR confront the issues related to the implementation, management, availability of human graders, and long-term financial sustainability. Hence, computer aided diagnosis tools are required for screening such a large population that require continuous follow-up for DR and to effectively facilitate in reducing the burden on the ophthalmologists (Jelinek and Cree, 2009; Walter et al., 2002). Such a tool would help clinicians in the identification, interpretation, and measurements of retinal abnormalities, and ultimately in the screening and monitoring of the disease.”
“Precise pixel-level annotations of lesions associated with DR such as MAs, HEs, SEs and EXs are invaluable resource for evaluating accuracy of individual lesion segmentation techniques. These precisely segmented lesions help in determining the disease severity and further act as a road-map that can assist to tap the progression of disease during follow-up procedures. Similarly, on the other hand, image-level expert labels for disease severity of DR, and DME are helpful in the development and evaluation of image analysis and retrieval algorithms.” (page no. 4 (line no. 38-45))

➢ Recent study highlighting importance of lesion segmentation is presented in the literature review:
“However, Lynch et al., 2017 demonstrated that the hybrid algorithms based on multiple semi-dependent CNNs might offer a more robust option for DR referral screening, stressing the importance of lesion segmentation.” (page no. 12 (line no. 270-273))

➢ Content highlighting importance of simultaneous DR and DME grading:
“Whereas, DR(n) and DME(n) are the predicted results, then correct instance is the case when the expert label for DR and DME matches with the predicted outcomes for both DR and DME. This was done since, even with presence of some exudation that may be categorized as mild DR, its location on the retina is also important governing factor (to check DME) to decide overall grade of disease. For instance, EXs presence in the macular region can affect vision of the patient to greater extent and hence, it should be dealt with priority for referral (that may otherwise be missed or cause delay in treatment with the present convention of only DR grading) in the automated screening.” (page no. 43 (line no. 1014-1022))

➢ Content presenting importance of OD and fovea detection:
“These techniques are shown to usually involve interdependence on the detection of anatomical structures (i.e. OD and fovea) with the lesion detection, and that in turn determines the automated DR screening outcome.” (page no. 8 (line no. 156-159))

“Localization and segmentation of OD and fovea facilitate the detection of retinal lesions as well as in the assessment (based on the geometric location of these lesions) of the severity and monitoring the progression of DR and DME.” (page no. 8 (line no. 160-162))

“There were three principal sub-challenges: lesion segmentation, disease severity grading, and localization and segmentation of retinal landmarks. These multiple tasks in IDRiD challenge allow to test the generalizability of the algorithms, and this is what makes it different from the existing ones.” (page no. 6 (line no. 88-91))

“This current progress in artificial intelligence provides an opportunity to the researchers for enhancing the performance of the DR referral system to more robust diagnosis system that can provide the quantitative information for multiple diseases matching the international standards of clinical relevance. Thus, this challenging design offers an avenue to gauge precise DR severity status and opportunity to deliver accurate measures for lesions, that could even help in the follow-up studies to observe changes in the retinal atlas.” (page no. 12 (line no. 277-283))

Hence, as the clinical relevance of this challenge is highlighted in the existing text, to avoid redundancy, we refrained from adding discussion about the same.
21. I believe it would be interesting to include some discussion regarding challenge organization. Since the different teams did not have much time to submit their methods and results, I strongly believe this had an influence on the results (and not only in the number of teams involved)

Response: We incorporated the opinion of participating teams for addressing this comment and have included discussion for the same in the manuscript.

➢ The following content (in blue) is added in the manuscript to address this comment:

“However, it seems there might be some impact of challenge duration, apart from the number of submissions, on the quality of developed solutions. Considering the time span from data availability to deadline of results submission, about one and a half month, was considerably tight for managing all tasks at the same time. For the team VRT who had been working on analyzing fundus images for more than a year when participated in the competition that attempting all tasks were possible, still it was challenging for them to commit all the tasks. However, it would be highly challenging for a newcomer to succeed in multiple tasks. In that sense, the competition period was not sufficient for perfecting all tasks. However, it would be enough for a competent participant, e.g. new entrants in the field as team SAIHST, to finish one task if the participant can focus on the competition completely.” (page no. 56 (line no. 1205-1216))

Reviewer #3:

Manuscript Rating Question(s): Scale [1-5]
The paper is of enough importance to warrant publication in MedIA 3
The paper is technically sound 4
The paper describes original work 3
The work is of interest to the MedIA audience 4
The paper contains material which might well be omitted 1
The paper makes adequate references? 3
The abstract is an adequate digest of the work reported 5
The introduction gives the background of the work 5
The summary and conclusions adequate 5
The authors explain clearly what they have done 5
The authors explain clearly why what they did was worth doing 5
The order of presentation is satisfactory 5
The English is satisfactory 5
If there are color figures included, are they helpful/necessary? 5
If there is a video, is it helpful/necessary? N/A

Comments

This paper describes the new IDRiD dataset and the challenge that was organized using it. Its contents are very clear and conveniently illustrated.
Public annotated datasets are a valuable scientific resource. Therefore, the presented work is commendable from this point of view. The quality of the annotations is an important point. As far as I have seen (without exhaustively reviewing the data) quality criteria are here met. The size of the dataset is another important criterion. In my opinion, this criterion is barely met for sub-challenges 1 and 3, that involve pixel-level segmentations. I am aware that manually producing segmentations - especially for lesions - is extremely time-consuming, but by current standards 81 images is really a small number. In the case of subchallenge 2, the lack of images seems evident, especially if you compare with a recent dataset as Kaggle's, of even an older one, as Messidor's. Moreover, from a real-world practical point of view, the fact that all images come from a single retinograph model, and are limited to good quality images, is a pity. In spite of these shortcomings, I believe that the dataset is a useful contribution to this domain.

Organizing challenges is also a useful contribution to the domain. However, the results seem similar to those of the state-of-the-art. In any case, they are not compared in any way with previous work. It would be interesting to at least recall the current state-of-the-art, even if it corresponds to other databases. In any case, no apparent break-through has been introduced by the winning solutions and as such are not very interesting from a scientific point of view.

Some minor remarks
---------------------

1. p. 3: citing (Abramoff et al., 2010) for sustaining the claim that "Early diagnosis and treatment of DR can prevent vision loss" is inadequate. Earlier reference would be more appropriate. I have the impression that other citations should also be checked from this point of view (like (Ting et al., 2016)).

   **Response:** We have thoroughly checked the manuscript and corrected these and other identified instances.

2. p. 11: 1,28,175 images?

   **Response:** We have corrected it to 128,175 in the manuscript.

3. section 5.2: "as follows:" - the algorithm has moved away.

   **Response:** We have corrected this in the modified manuscript. The change could be observed on page no. 43.

4. Fig. 5(a): the ROC curves should not go under the diagonal. this is not really a minor remark, by the way.

   **Response:** We would like to humbly bring into notice that the evaluation measure used for determining the lesion segmentation accuracy was AUPR curves, where the area under the curve may go below 50%.

5. It would be interesting to comment on the late participation of the CBER team. Why have their results been included? Is it because they obtain interesting scores, without using deep learning?

   **Response:** We have addressed this comment in section 4. Challenge organization, just before the start of details regarding phase 4.

   ➢ The following content is added in the manuscript to address this comment:

   “Amongst invited, 13 teams confirmed their participation in the on-site challenge, whereas, two teams declined to participate due to other commitments and one team was not able arrange financial support in the limited time.” (page no. 18-19. (line no. 398-401))
Conflict of Interest Statement

To,
The Editor-in-chief,
Medical Image Analysis, Elsevier Journal.

We would like to submit our challenge summary paper entitled “IDRiD: Diabetic Retinopathy - Segmentation and Grading Challenge” authored by Prasanna Porwal, Samiksha Pachade, Manesh Kokare, Girish Deshmukh, Jaemin Son, Woong Bae, Lihong Liu, Jianzong Wang, Xinhui Liu, Liangxin Gao, TianBo Wu, Jing Xiao, Fengyan Wang, Baocai Yin, Yunzhi Wang, Gopichandh Danala, Linsheng He, Yoon Ho Choi, Yeong Chan Lee, Sang-Hyuk Jung, Zhongyu Li, Xiaodan Sui, Junyan Wu, Xiaolong Li, Ting Zhou, Janos Toth, Agnes Baran, Avinash Kori, Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Xingzheng Lyu, Li Cheng, Qinhao Chu, Pengcheng Li, Xin Ji, Sanyuan Zhang, Yaxin Shen, Ling Dai, Oindrila Saha, Rachana Sathish, Tânia Melo, Teresa Araújo, Balazs Harangi, Bin Sheng, Ruogu Fang, Debdoott Sheet, Andras Hajdu, Yuanjie Zheng, Ana Maria Mendonça, Shaoting Zhang, Aurélio Campilho, Bin Zeng, Dinggang Shen, Luca Giancardo, Gwenolé Quellec, and Fabrice Meriaudeau and also to be considered for publication as a research paper in Medical Image Analysis, Elsevier Journal.

We confirm that this manuscript has not been submitted for consideration by another journal. All authors have approved the manuscript and agree with submission to Medical Image Analysis, Elsevier Journal.

The authors have no conflicts of interest to declare.

Best Regards,
Prasanna Porwal (porwalprasanna@sggs.ac.in)