

2018; Ker et al., 2018). Among various methodological variants of deep learning, Convolutional Neural Networks (CNNs or ConvNets) are most popular in 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 retinal image analysis because of its unique characteristic of preserving local image relations. Majority of approaches in literature employ deep learning to retinal images by utilizing “off-the-shelf CNN” features as complementary information channels to other hand-crafted 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 detection of DR (Rangrej and Sivaswamy, 2017). The authors (Fu et al., 2016) employ fully connected conditional random fields along with CNN to integrate 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 of 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). García et al. (2017) trained the “CNN from scratch” and compared it with fine-tuning results based on two other existing architectures. Recently, Shah et al. (2018) demonstrated that 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 vasculature embedding that leverages internal representation of a new encoder-enhanced CNN, demonstrating improvement in 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) was proposed for DR screening and trained using 75,137 images obtained from EyePACS system (Cuadros and Bresnick, 2009), where an additional classifier was further employed on the CNN-derived features to determine if an 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 labels. There are some hybrid algorithms, in which multiple, semi-dependent CNN's are trained based on the appearance of retinal lesions (Abràmoff et al., 2016; Quellec et al., 2016). A step further, researchers (Quellec et al., 2017) demonstrated an ability of lesion segmentation based on CNN trained for image-level classification. However, Lynch et al. (2017) demonstrated that 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 researchers for enhancing the performance of DR referral system to a more robust diagnosis system that can provide quantitative information for multiple diseases matching international standards of clinical relevance. Thus, the presented challenge design offers an avenue to gauge precise DR severity status and op-

portunity to deliver accurate measures for lesions, that could even help in follow-up studies to observe changes in the retinal atlas.

3. Indian diabetic retinopathy image dataset

The IDRiD dataset (Porwal et al., 2018a) was created from real clinical exams acquired at an eye clinic located in Nanded, (M.S.), India. Retinal photographs of people affected by diabetes were captured with focus on macula using Kowa VX – 10 α fundus camera. Prior to image acquisition, pupils of all subjects were dilated with one drop of tropicamide at 0.5% concentration. The captured images have 50° field of view, resolution of 4288 \times 2848 pixels and are 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 DR 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 Annotation*: This type of annotation is useful in 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 were annotated at the 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 the pixel-level and binary masks for all 81 images are provided in the same format. All of these annotations play a vital role in research for computational analysis of segmenting lesions within the image.

2. *Image Level Grading*: It consists of information meant to describe the overall risk factor associated with an entire image. Two medical experts provided adjudicated consensus grades to the full set of 516 images with a variety of pathological conditions of DR and DME. Grading for all images is available in the CSV file. The diabetic retinal images were classified into separate groups according to the International Clinical Diabetic Retinopathy Scale (Wu et al., 2013), confined to image under observation, as shown in Table 1.

The DME severity was decided based on occurrences of EXs near to macula center region (Decencièrre et al., 2014) as shown in Table 2.

3. *OD and Fovea center co-ordinates* The OD and fovea center locations are marked for all 516 images and the markup is available as a separate CSV file.

Table 1

DR Severity Grading. NPDR: Non-proliferative DR and PDR: Proliferative DR

DR Grade	Findings
0: No apparent retinopathy	No visible sign of abnormalities
1: Mild NPDR	Presence of MAs only
2: Moderate NPDR	More than just MAs but less than severe NPDR
3: Severe NPDR	Any of the following: <ul style="list-style-type: none"> • >20 intraretinal HES • Venous beading • Intraretinal microvascular abnormalities • no signs of PDR
4: PDR	Either or both of the following: <ul style="list-style-type: none"> Neovascularization Vitreous/pre-retinal HE

Table 2

Risk of DME.

DME Grade	Findings
0	No Apparent EX(s)
1	Presence of EX(s) outside the radius of one disc diameter from the macula center
2	Presence of EX(s) within the radius of one disc diameter from the macula center

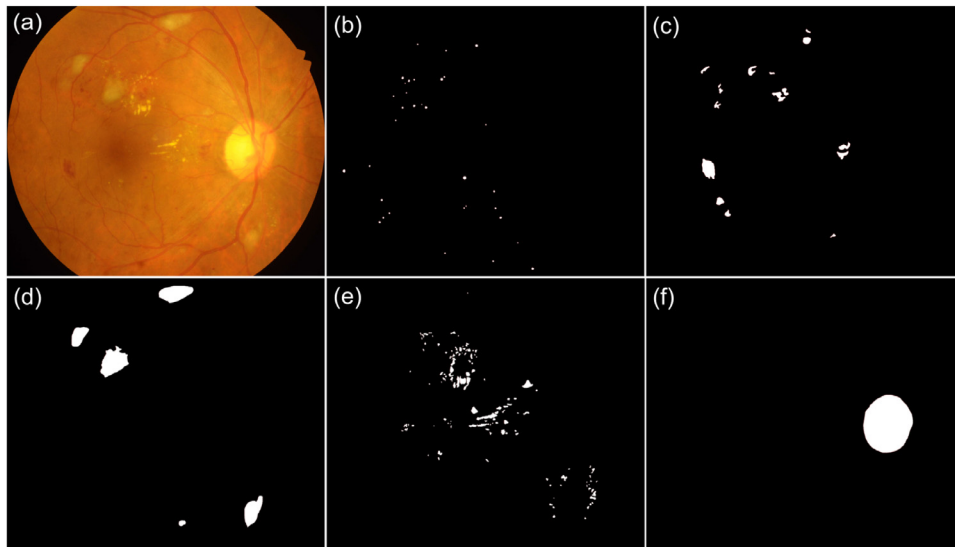


Fig. 2. Retinal photograph and different pixel-level annotations: (a) sample fundus image from the IDRiD dataset; sample ground truths for (b-f) MAs, HES, SEs, EXs and OD respectively.

The IDRiD dataset is available from IEEE Dataport Repository⁴ under a Creative Commons Attribution 4.0 license. More detailed information about the data is available in the data descriptor (Porwal et al., 2018b). Tables A.1 and A.2 highlight a comparative strength of this dataset with respect to existing datasets. IDRiD is the only dataset that provides all three types of annotations mentioned above. This streamlined collection of annotations would allow it to be utilized in research and lead to the development of better generalizable models for image analysis, enabling further progress in 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 success of the contest. Fig. 3 depicts the work-flow of the overall challenge organization. The challenge was officially announced at the ISBI - 2018 website⁵ on 15th October 2017.

The challenge was subdivided into three sub-challenges as follows:

1. Lesion Segmentation: Segmentation of retinal lesions associated with DR such 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 ground truths was separated into a training set and test set. For images with pixel-level annotations, data was separated as 2/3 for training (Set-A) and 1/3 for testing (Set-B) (See Table 3).

Table 3

Stratification of retinal images annotated at pixel level for different types of retinal lesions.

Lesion Type	Set - A Images	Set - B Images
MA	54	27
HE	53	27
SE	26	14
EX	54	27

Table 4

Stratification of retinal images graded for DR and DME.

DR Grade	Set-A	Set-B	DM Grade	Set-A	Set-B
0	134	34	0	177	45
1	20	5	1	41	10
2	136	32	2	195	48
3	74	19			
4	49	13			

Similarly, data for OD segmentation (part of sub-challenge – 3) was divided in the same ratio into Set-A (54 images) and Set-B (27 images). Since the output of algorithms would be representative of learned perceptive patterns. The data for lesion and OD segmentation tasks were carefully split in such a way that it provides enough representative data to be learned and a holdout proportion that could be later used to gauge the algorithm performance. 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) was supported by the research outcome (Dobbins 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 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 to the DR and DME grades before splitting. A breakdown of the details of the dataset is shown in Table 4.

⁴ <https://ieee-dataport.org/open-access/indian-diabetic-retinopathy-image-dataset-idrid>

⁵ <https://biomedicalimaging.org/2018/challenges/>

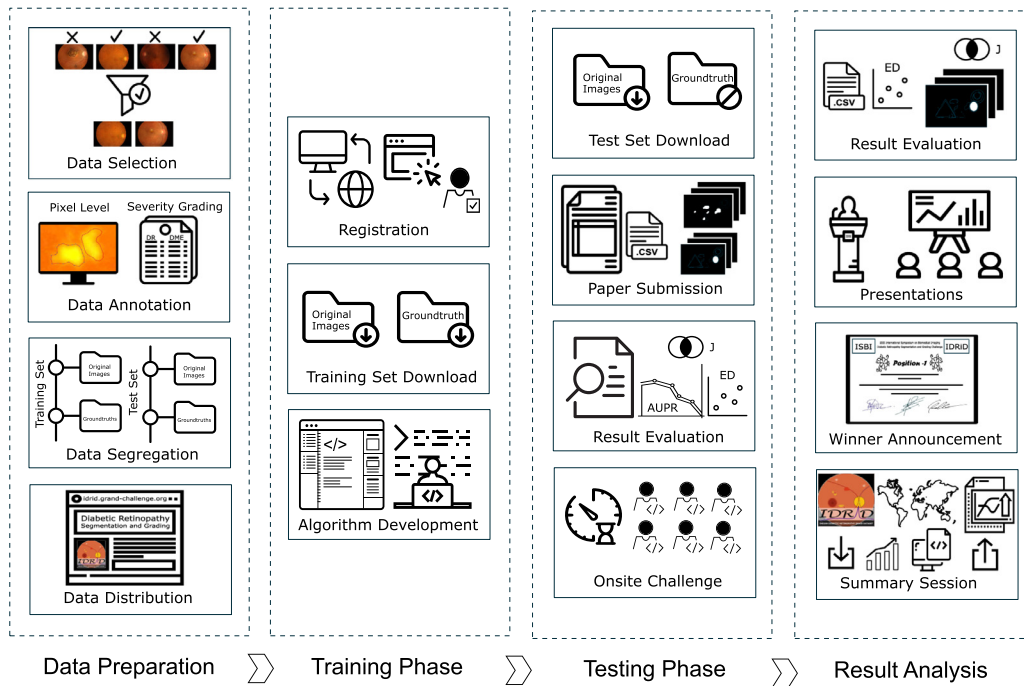


Fig. 3. Workflow of the ISBI - 2018: Diabetic Retinopathy – Segmentation and Grading Challenge.

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

Stage-2. Registration and release of the training data: Registration of challenge for consideration to ISBI on-site contest was open from the launch of the grand-challenge website (i.e. 25th October 2017) till the deadline for submission of results (i.e. 11th March 2018). Interested research teams could register through challenge website for one or all sub-challenges. The first part of data, i.e., Set-A (images and ground truths) was made available to participants of challenge on 20th 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 a condition that external datasets should be publicly available.

Stage-3. Release of test data: Set-B (only images) for sub-challenge – 1 was released on 20th February, 2018. For other two sub-challenges, Set-B was released on 4th April which was part of 'on-site' challenge. Organizers refrained from an on-site evaluation of sub-challenge – 1 considering timing constraints in the evaluation of results for image segmentation tasks.

Submissions were sought for either of the following 8 different tasks corresponding to 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: OD 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 to the organizers till 11th March. 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, 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 challenge organizers and their performance was displayed on the leaderboard of the challenge website. For sub-challenge – 1, teams were assessed based on the performance of results submitted on a test set, whereas for other two sub-challenges assessment was done using results on a training set obtained through leave one out cross-validation approach. In this phase, it received a very good response from the research community with 148 submissions by 37 different teams, out of which 16 teams were shortlisted for participation to 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 to 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 the 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,

⁶ <https://grand-challenge.org/>

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), hemorrhages (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 (SC3) 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 the text as LzyUNCC-I) and LzyUNCC_Fusion (renamed in the 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.

Team Name	Authors	DL	SC1				SC2	SC3		
			MA	HE	SE	EX		ODD	FD	ODS
 VRT	Jaemin Son et al.	✓	✓	✓	✓	✓	✓	✓	✓	✓
 iFLYTEK-MIG	Fengyan Wang et al.	✓	✓	✓	✓	✓	×	×	×	×
 PATech	Liu Lihong et al.	✓	✓	✓	×	✓	×	×	×	×
 SOONER	Yunzhi Wang et al.	✓	✓	✓	✓	✓	×	×	×	×
 SAIHST	Yoon Ho Choi et al.	✓	×	×	×	✓	×	×	×	×
 LzyUNCC	Zhongyu Li et al.	✓	×	×	✓	✓	✓	×	×	×
 SDNU	Xiaodan Sui et al.	✓	✓	✓	✓	✓	×	✓	✓	✓
 Mammoth	Junyan Wu et al.	✓	×	×	×	×	✓	×	×	×
 HarangiM1-M2	Balazs Harangi et al.	✓	×	×	×	×	✓	×	×	×
 AVSASVA	Varghese Alex et al.	✓	×	×	×	×	✓	×	×	×
 DeepDR	Ling Dai et al.	✓	×	×	×	×	×	✓	✓	×
 ZJU-BII-SGEX	Xingzheng Lyu et al.	✓	×	×	×	×	×	✓	✓	✓
 IITkgpKLIV	Oindrila Saha et al.	✓	×	×	×	×	×	×	×	✓
 *CBER	Ana Mendonça et al.	×	×	×	×	×	×	✓	✓	✓

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 the 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 results 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 was 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 Table 5 and the 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 to CNN and its variants utilized by participating teams. Further, it summarizes the solutions and related technical specifications. For the detailed description of a particular approach, please refer to proceedings of ISBI Grand Challenge Workshop at https://idrid.grand-challenge.org/Challenge_Proceedings/.

For the input image, CNN transforms raw image pixels on one end to generate a single differentiable score function at the other end. It exploits three mechanisms – sparse connections (*a.k.a.* local receptive field), weight sharing and invariant (or equivariant) rep-

resentation – 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:

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 ℓ^{th} layer, the feature value in i^{th} feature map, $A_{x,y,i}^\ell$, is obtained as:

$$A_{x,y,i}^\ell = f((w_i^\ell)^T I_{x,y}^\ell + b_i^\ell) = f(C_{x,y,i}^\ell) \quad (1)$$

Where the parameters w_i^ℓ and b_i^ℓ are weight vector and bias term of i^{th} filter of ℓ^{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^ℓ that generates the feature map $C_{x,y,i}^\ell$ is shared, reducing model complexity and making network easier to train.

POOL layer aims to achieve translation-invariance by reducing the resolution of feature maps. Each unit in a feature map of POOL layer is derived using a subset of units within sparse connections from a 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 initial CONV layers detect low-level features such as edges and curves, while the kernels in higher layers are learned to encode more abstract features. The sequence of several CONV and POOL layers gradually extract higher-level feature representation.