

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).

















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

Table 9

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

















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

Table 10

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

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

OD and Fovea detection tasks. But the winning entries for OD segmentation task were from teams ZJU-BII-SGEX, VRT and IITKgp-KLIV. Some sample OD segmentation results from these teams are illustrated in Fig. 9.

Fig. 10 shows box-plots illustrating the range of Euclidean distances from the center of (a) OD and (b) fovea as well as (c) spread of Jaccard index for OD segmentation.

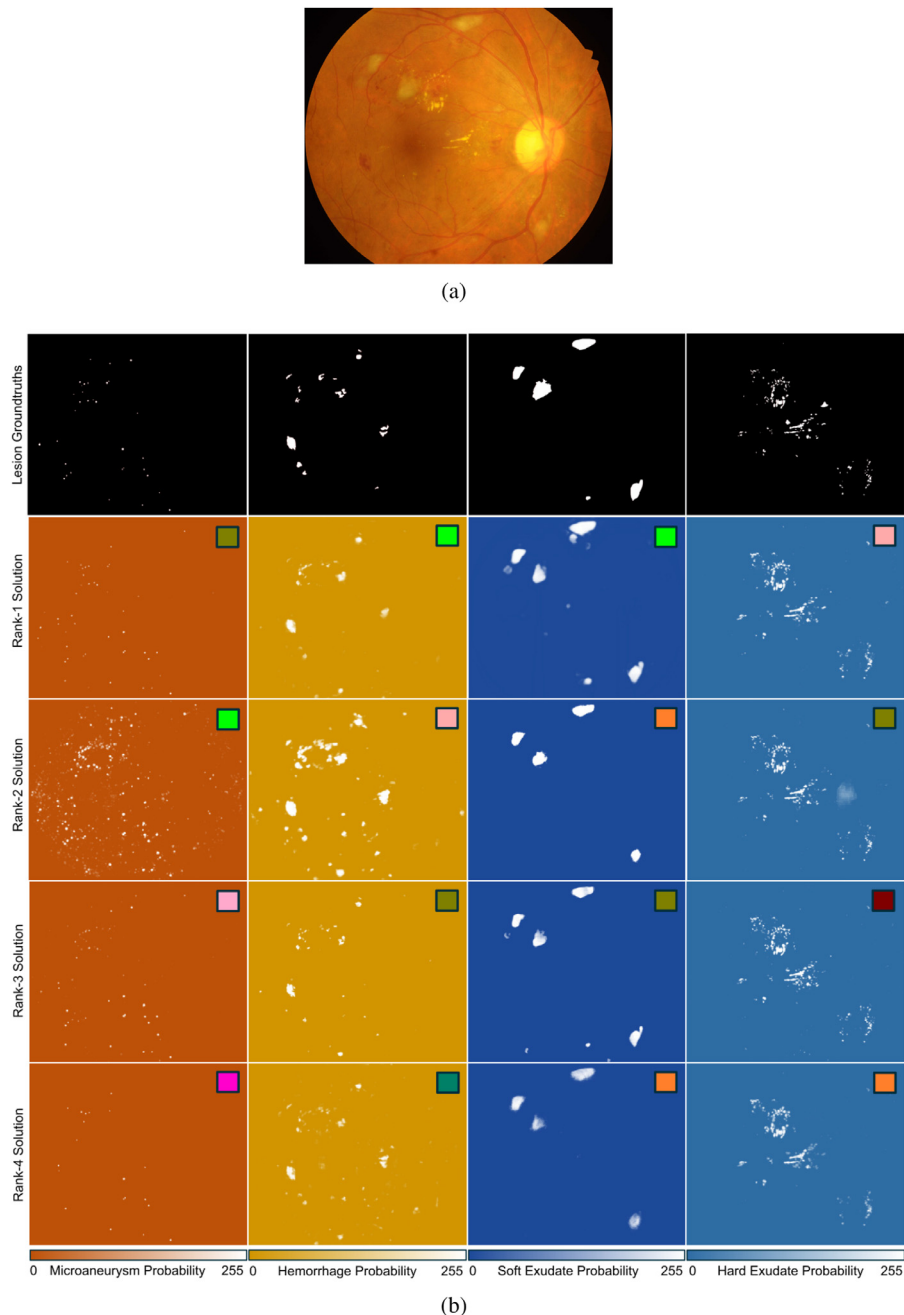


Fig. 6. Illustration of lesion segmentation results: (a) sample image and (b) segmentation outcome of top-4 teams (from left to right) (i) MAs, (ii) HEs, (iii) SEs, and (iv) EXs in retinal fundus images. Top row corresponds to ground truths, second row to entry from top performing team, similarly, third, 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.

Table 11

"On-site" leaderboard highlighting performance of top 5 teams in OD segmentation task on the test dataset. It details the approach followed by respective team and external dataset used for training their model (if any). J: Jaccard index.

Team name	J	Rank	Approach	Input size (Pixels)	External dataset
ZJU-BII-SGEX	0.9338	1	Mask R-CNN	1024 × 1024	RIGA
VRT	0.9305	2	U-Net	640 × 640	DRIVE, DRIONS-DB
IITKgpKLIV	0.8572	3	SegNet	536 × 356	Drishti-GS
SDNU	0.7892	4	Mask R-CNN	1984 × 1318	–
CBER	0.8912	–	Handcrafted Features	536 × 356	–

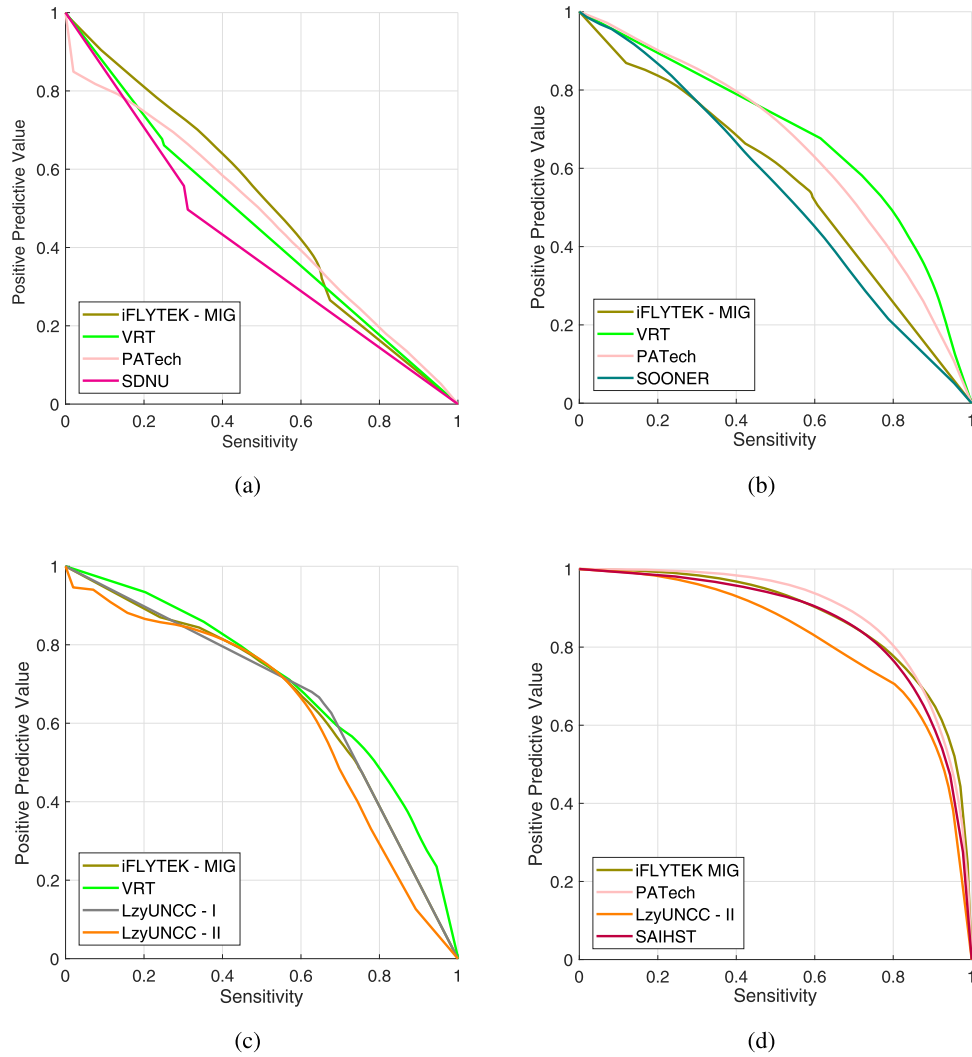


Fig. 7. The AUPR curves for the four top performing individual methods on the test dataset. These curves plot the sensitivity versus the positive predictive values for the different lesions, namely, (a) MAs, (b) HEs, (c) SEs, and (d) EXs.

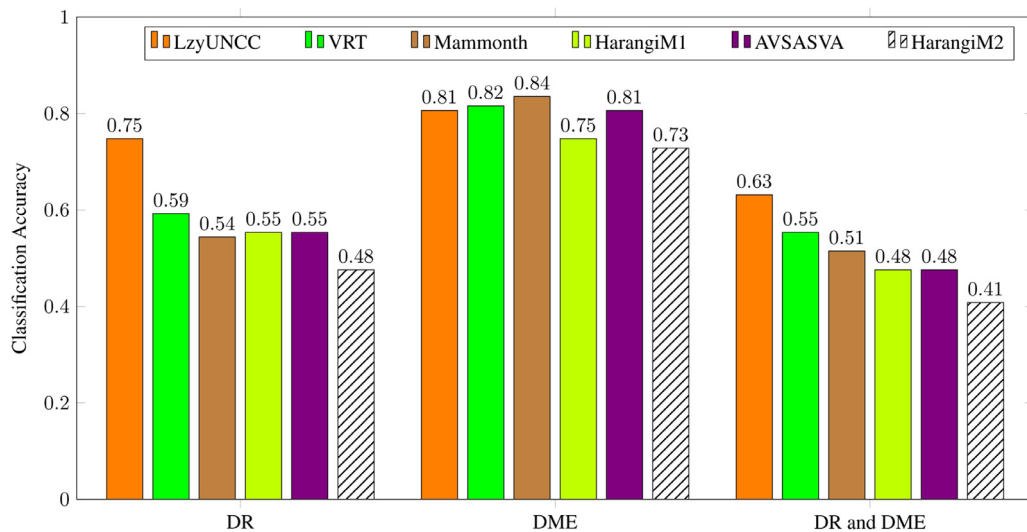


Fig. 8. Barplots showing separate and simultaneous classification accuracy of solutions developed by top-6 teams for grading of DR and DME.

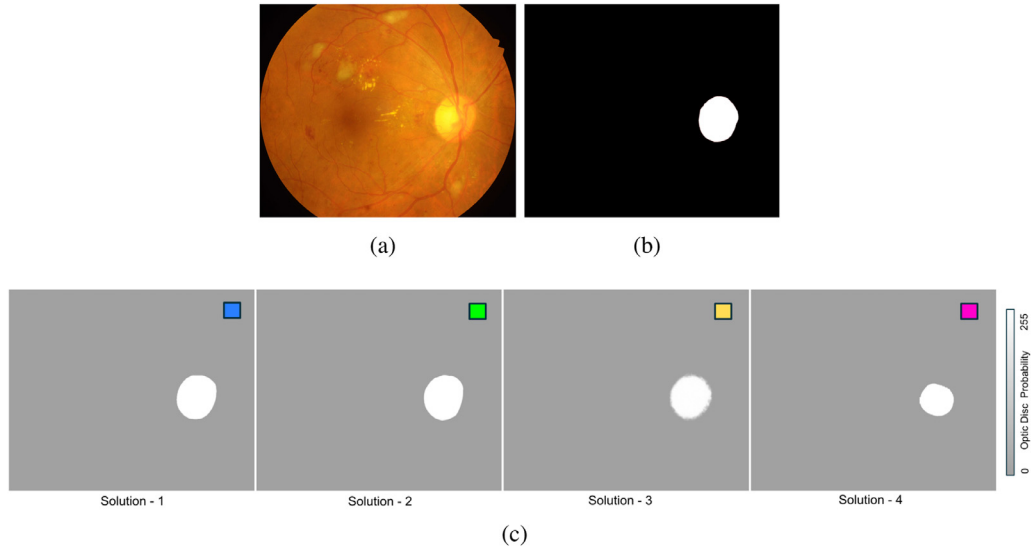


Fig. 9. Illustration of OD segmentation results: (a) sample image, (b) OD ground truth and (c) segmentation outcome of top-4 teams (from left to right).

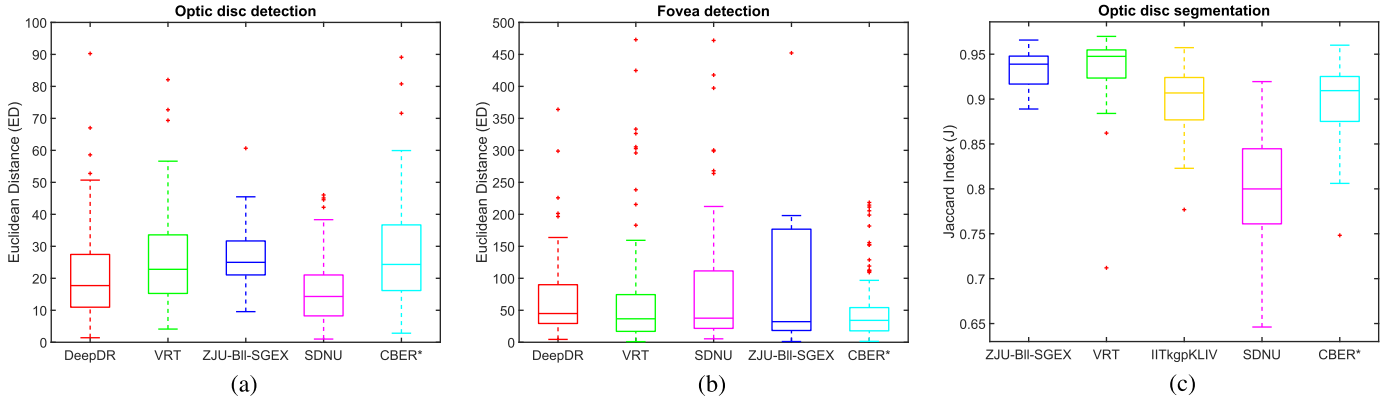


Fig. 10. Boxplots (a,b) showing dispersion of Euclidean distance for individual methods for OD and fovea and (c) showing the dispersion of Jaccard index for OD segmentation task. Boxplots show quartile ranges of the scores on the test dataset; plus sign indicate outliers (full range of data is not shown).

8. Discussion and conclusion

In this paper, we have presented the details of IDRid challenge including 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 localization 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 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). When the performance of top solutions was analyzed by computing the area under ROC curve (AUC) at the pixel level, in threshold range [0:0.01:1], it resulted in AUC of 0.8263, 0.9716, 0.9540 and 0.9883 for MA, HE, SE and EX respectively. The best approach for lesion segmentation used U-Net, with data augmentation and ad-

dition of dense block to extract features efficiently, boosting results significantly. Fig. 11 highlights the performance of top solution for EX that performs significantly well in the presence of normal retinal structures and different challenging circumstances.

From the top-performing approaches, it is evident that solving the data imbalance problem improves the model performance significantly. Since background overwhelms foreground i.e. there are more background pixels than lesion pixels (see Fig. 6), the loss during training is more effectively back-propagated than that of the foreground that penalizes false negatives, boosting the sensitivity of lesion segmentation. In general, the architectural modifications to U-Net-based networks provided widely varying results for the different types of lesion.

For instance, the cascaded CNN approach yielded the best score for MAs segmentation, as it adds modules to reduce false positives. This approach dramatically impacts MA segmentation performance due to the class imbalance of the task. Further, Fig. 12 shows that some false positives detected by 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 accuracy ranged between 0.4078 and