

Manuscript Number: MEDIA-D-19-00049R1

Title: IDRiD: Diabetic Retinopathy - Segmentation and Grading Challenge

Article Type: Research Paper

Keywords: Diabetic Retinopathy; Retinal image analysis; Deep learning;
Challenge

Corresponding Author: Mr. Prasanna Porwal, M.Tech

Corresponding Author's Institution: Shri Guru Gobind Singhji Institute of
Engineering and Technology

First Author: Prasanna Porwal, M.Tech

Order of Authors: Prasanna Porwal, M.Tech; Samiksha Pachade, M.Tech;
Manesh Kokare, PhD; Girish Deshmukh, M.S.; 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; Qin hao Chu; Pengcheng Li; Xin Ji; Sanyuan Zhang;
Yaxin Shen; Ling Dai; Oindrila Saha; Rachana Sathish; Tânia Melo; Teresa
Araújo; Balazs Harangi, PhD; Bin Sheng, PhD; Ruogu Fang, PhD; Debdoot
Sheet, PhD; Andras Hajdu, PhD; Yuanjie Zheng, PhD; Ana Mendonça, PhD;
Shaoting Zhang, PhD; Aurélio Campilho, PhD; Bin Zeng, PhD; Dinggang Shen,
PhD; Luca Giancardo, PhD; Gwenolé Quellec, PhD; Fabrice Meriaudeau, PhD

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://iee-dataport.org/open-access/indian-diabetic-retinopathy-image-dataset-idrid>

IDRiD: Diabetic Retinopathy - Segmentation and Grading Challenge

Prasanna Porwal^{a,b,1,*}, Samiksha Pachade^{a,b,1}, Manesh Kokare^{a,1}, Girish Deshmukh^{c,1}, Jaemin Son^d, Woong Bae^d, Lihong Liu^e, Jianzong Wang^e, Xinhui Liu^e, Liangxin Gao^e, TianBo Wu^e, Jing Xiao^e, Fengyan Wang^f, Baocai Yin^f, Yunzhi Wang^g, Gopichandh Danala^g, Linsheng He^g, Yoon Ho Choi^h, Yeong Chan Lee^h, Sang Hyuk Jung^h, Zhongyu Liⁱ, Xiaodan Sui^j, Junyan Wu^l, Xiaolong Li^m, Ting Zhouⁿ, János Tóth^o, Agnes Baran^o, Avinash Kori^p, Varghese Alex^p, Sai Saketh Chennamsetty^p, Mohammed Safwan^p, Xingzheng Lyu^{q,r}, Li Cheng^f, Qin hao Chu^s, Pengcheng Li^s, Xin Ji^l, Sanyuan Zhang^q, Yaxin Shen^{u,v}, Ling Dai^{u,v}, Oindrila Saha^x, Rachana Sathish^x, Tânia Melo^y, Teresa Araújo^{y,z}, Balázs Harangi^o, Bin Sheng^{u,v}, Ruogu Fang^w, Debdoot Sheet^x, Andras Hajdu^o, Yuanjie Zheng^j, Ana Maria Mendonça^{y,z}, Shaoting Zhangⁱ, Aurélio Campilho^{y,z}, Bin Zheng^g, Dinggang Shen^k, Luca Giancardo^{b,1}, Gwenolé Quéllec^{aa,1}, Fabrice Mériaudeau^{ab,ac,1}

^aShri 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

^fFLYTEK Research, Hefei, China

^gSchool of Electrical and Computer Engineering, University of Oklahoma, USA

^hSamsung Advanced Institute for Health Sciences & Technology (SAIHST), 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 Shangong 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

^{aa}INSERM, UMR 1101, Brest, France

^{ab}Department of Electrical and Electronic Engineering, Universiti Teknologi PETRONAS, Malaysia

^{ac}ImViA/IPTIM, Université de Bourgogne, Dijon, France

*Corresponding author

Email address: porwal.prasanna@sngs.ac.in (Prasanna Porwal)

¹These authors co-organized the challenge. All others contributed results of their algorithm(s) presented in the paper

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 **1. Introduction**

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

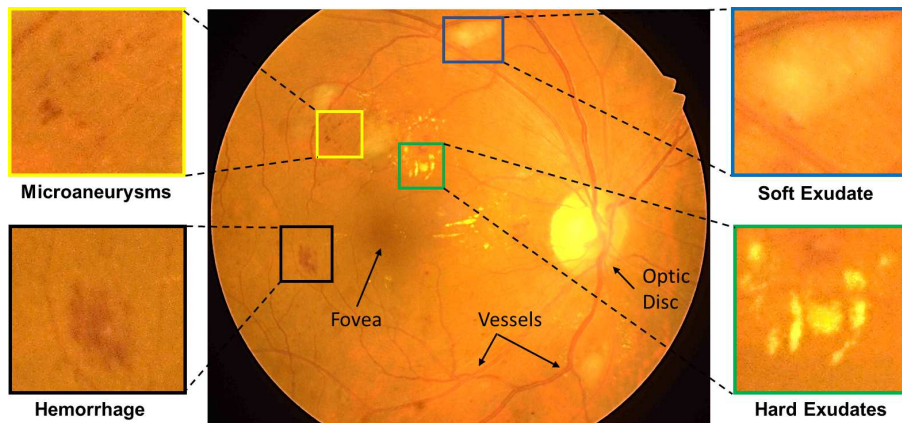


Fig. 1. Illustration of retinal image (in center) by highlighting normal structures (blood vessels, optic disc and fovea center) and abnormalities associated with DR: Enlarged regions (in left) MAs, and HEs and (in right) SEs, and EXs.

17 Early diagnosis and treatment of DR can prevent vision loss. Hence, diabetic pa-
18 tients are typically referred for retinal screening once or twice a year (Ferris, 1993;

19 Kollias and Ulbig, 2010; Ting et al., 2016). The diabetic eye care is mainly reliant
20 on the number of ophthalmologists and necessary health care infrastructure (Jones and
21 Edwards, 2010; Lin et al., 2016). In the Indian subcontinent, ophthalmologist to popu-
22 lation ratio is 1:107,000, however, in urban regions this ratio is 1:9000 whereas in rural
23 parts there is only one ophthalmologist for 608,000 inhabitants (Raman et al., 2016).
24 By 2045, India alone is projected to have approximately 151 million people with dia-
25 betes and one-third of them are expected to have DR (Atlas, 2017). Programs to screen
26 such a large population for DR confront the issues related to the implementation, man-
27 agement, availability of human graders, and long-term financial sustainability. Hence,
28 computer aided diagnosis tools are required for screening such a large population that
29 require continuous follow-up for DR and to effectively facilitate in reducing the bur-
30 den on the ophthalmologists (Jelinek and Cree, 2009; Walter et al., 2002). Such a tool
31 would help clinicians in the identification, interpretation, and measurements of retinal
32 abnormalities, and ultimately in the screening and monitoring of the disease. The recent
33 scientific advances in computing capacity and machine learning approaches provide an
34 avenue to the biomedical scientists to meet the desideratum of clinical practice (Short-
35 liffe and Blois, 2006; Patton et al., 2006). To meet this need raw images along with
36 precise pixel or image level expert annotations (also known as ground truths) play an
37 important role to facilitate the research community for the development, validation,
38 and comparison of DR lesion segmentation techniques (Trucco et al., 2013). Precise
39 pixel-level annotations of lesions associated with DR such as MAs, HEs, SEs and EXs
40 are invaluable resource for evaluating accuracy of individual lesion segmentation tech-
41 niques. These precisely segmented lesions help in determining the disease severity
42 and further act as a road-map that can assist to tap the progression of disease during
43 follow-up procedures. Similarly, on the other hand, image-level expert labels for dis-
44 ease severity of DR, and DME are helpful in the development and evaluation of image
45 analysis and retrieval algorithms. This necessity has led several research groups to
46 develop and share retinal image datasets, namely Messidor (Decencière et al., 2014),
47 Kaggle (Cuadros and Bresnick, 2009), ROC (Niemeijer et al., 2010), E-Ophtha (De-
48 cencière et al., 2013), DiaretDB (Kauppi et al., 2012), DRIVE (Staal et al., 2004),
49 STARE (Hoover, 1975), ARIA (Farnell et al., 2008) and HEI-MED (Giancardo et al.,

50 2012).

51 Further, two challenges were organized in the context of DR, namely Retinopathy
52 Online Challenge (ROC)² and Kaggle DR detection challenge³. ROC was organized
53 with the goal of detecting MAs. Whereas, the Kaggle challenge aimed to get solution
54 for determining the severity level of DR. These challenges enabled advances in the field
55 by promoting the participation of scientific research community from all over the globe
56 on a competitive at the same time constructive setting for scientific advancement. Pre-
57 vious efforts have made good progress using image classification, pattern recognition,
58 and machine learning. The progress through last two decades has been systematically
59 reviewed by several research groups (Patton et al., 2006; Winder et al., 2009; Abràmoff
60 et al., 2010; Mookiah et al., 2013a; Jordan et al., 2017; Nørgaard and Grauslund, 2018).

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

75 In order to address these issues, we introduced a new dataset called Indian Diabetic
76 Retinopathy Image Dataset (IDRiD) (Porwal et al., 2018a). Further, it was used as a
77 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>

78 mentation and Grading” in conjunction with ISBI - 2018. The IDRiD dataset provides
79 expert markups of typical DR lesions and normal retinal structures. It also provides
80 disease severity level of DR, and DME for each image in the database. This challenge
81 brought together the computer vision and biomedical researchers with an ultimate aim
82 to further stimulate and promote research, as well as to provide a unique platform for
83 the development of a practical software tool that will support efficient and accurate
84 measurement and analysis of retinal images that could be useful in DR management.
85 Initially, a training dataset along with the ground truth was provided to participants for
86 the development of their algorithms. Later, the results were judged on the performance
87 of these algorithms on test dataset. Success was measured by how closely the algo-
88 rithmic outcome matched the ground truth. There were three principal sub-challenges:
89 lesion segmentation, disease severity grading, and localization and segmentation of
90 retinal landmarks. These multiple tasks in IDRiD challenge allow to test the general-
91 ization of the algorithms, and this is what makes it different from the existing ones.
92 Further, this challenge seeks an automated solution to predict the severity of DR and
93 DME simultaneously. It was projected as an individual task to increase the difficulty
94 level of this challenge as compared to the Kaggle DR challenge i.e. for a given image,
95 the predicted severity for both DR and DME should be correct to count for scoring the
96 task.

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

107 **2. Review of Retinal Image Analysis for the detection of DR**

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

117 *2.1. Non-deep learning methods*

118 The general framework for retinal image analysis through traditional handcrafted
119 features based approaches involve several stages, typically: a preprocessing stage for
120 contrast enhancement or non-uniformity equalization, image segmentation, feature ex-
121 traction, and classification. The feature extraction strategy varies according to the ob-
122 jective involved i.e. retinal lesion detection, disease screening or landmark localization.
123 In 2006, one research group (Patton et al., 2006) outlined the principles upon which
124 retinal image analysis is based and discussed the initial techniques used to detect the
125 retinal landmarks and lesions associated with DR. Later, one another group (Winder
126 et al., 2009) reported an analysis of the work in the automated analysis of DR dur-
127 ing 1998–2008. They categorized the literature into a series of operations or steps as
128 preprocessing, vasculature segmentation, localization, and segmentation of the optic
129 disk (OD), localization of the macula and fovea, detection and segmentation of le-
130 sions. Some of the review articles (Abràmoff et al., 2010; Jordan et al., 2017) provide
131 a brief introduction to quantitative methods for the analysis of fundus images with
132 a focus on identification of retinal lesions and automated techniques for large scale
133 screening for retinal diseases. Majority of attempts in the literature are towards exclu-
134 sive detection and/or segmentation of one type of lesions (either MAs, HEs, EXs or
135 SEs) from an image. Some of the common approaches involved for lesion segmen-
136 tation are mathematical morphology (Joshi and Karule, 2019; Hatanaka et al., 2008;

137 Zhang et al., 2014), region growing (Fleming et al., 2006; Li and Chutatape, 2004),
138 and supervised (Wu et al., 2017; Zhou et al., 2017; Garcia et al., 2009; Tang et al.,
139 2013). Apart from these approaches, in case of MAs, most initial studies shown the
140 effectiveness of template matching (Quellec et al., 2008), entropy thresholding (Das
141 et al., 2015), radon space (Giancardo et al., 2011), sparse representation (Zhang et al.,
142 2012; Javidi et al., 2017), hessian based region descriptors Adal et al. (2014), dictio-
143 nary learning (Rocha et al., 2012). On the other hand, for exclusive segmentation of
144 HEs, super-pixel based features (Tang et al., 2013; Romero-Oraá et al., 2019) were
145 found to be effective. These red lesions (both MAs and HEs) are also frequently
146 detected together using dynamic shape features (Seoud et al., 2016), filter response
147 and multiple kernel learning (Srivastava et al., 2017) and hybrid feature extraction ap-
148 proach (Niemeijer et al., 2005). Similarly, for EXs researchers relied on approaches
149 like clustering (Osareh et al., 2009), model-based (Sánchez et al., 2009; Harangi and
150 Hajdu, 2014), ant colony optimization (ACO) (Pereira et al., 2015) and contextual in-
151 formation (Sánchez et al., 2012). Whereas, for SEs researchers utilized Scale Invariant
152 Feature Transform (SIFT) (Naqvi et al., 2018), adaptive thresholding and ACO (Sreng
153 et al., 2019). Further, several approaches were devised for multiple lesion detection
154 such as multiscale amplitude-modulation-frequency-modulation (Agurto et al., 2010),
155 machine learning (Roychowdhury et al., 2014), a combination of Hessian multiscale
156 analysis, variational segmentation and texture features (Figueiredo et al., 2015). These
157 techniques are shown to usually involve interdependence on the detection of anatomi-
158 cal structures (i.e. OD and fovea) with the lesion detection, and that in turn determines
159 the automated DR screening outcome.

160 Localization and segmentation of OD and fovea facilitate the detection of retinal
161 lesions as well as in the assessment (based on the geometric location of these lesions)
162 of the severity and monitoring the progression of DR and DME. Hence, several ap-
163 proaches have been proposed for localization of OD, most of them utilized the OD
164 properties like intensity, shape, color, texture, etc. and many others showed the ef-
165 fectiveness of mathematical morphology (Morales et al., 2013; Marin et al., 2015),
166 template matching (Giachetti et al., 2014), deformable models (Yu et al., 2012; Wu
167 et al., 2016) and intensity profile analysis (Kamble et al., 2017; Uribe-Valencia and

168 Martínez-Carballido, 2019). Further, the approaches utilized for OD segmentation
169 are based on level set (Yu et al., 2012), thresholding (Marin et al., 2015), active con-
170 tour (Mary et al., 2015) and shape modeling (Cheng et al., 2015), clustering (Thakur
171 and Juneja, 2017), and hybrid (Bai et al., 2014) approaches. Similarly, the fovea is de-
172 tected mostly using the geometric relationship with OD and vessels through morpho-
173 logical (Welfer et al., 2011), thresholding (Gegundez-Arias et al., 2013), template (Kao
174 et al., 2014) and intensity profile analysis (Kamble et al., 2017) techniques. Poor per-
175 formance on detection of the normal anatomical structures could adversely affect lesion
176 detection and screening accuracy. For instance, consider the mathematical morphol-
177 ogy based techniques presented in 2002 (Walter et al., 2002), 2008 (Sopharak et al.,
178 2008) and 2014 (Zhang et al., 2014). These works demonstrate how the morphological
179 processing-based approaches evolved by including multiple steps for the final objective
180 of exudate detection. In the initial efforts, Walter et al. devised a technique for OD and
181 EXs segmentation, later removed the OD to obtain the exudate candidates. Similarly,
182 Sopharak et al. achieved the same objective with the detection, and removal of OD
183 and vessels. Recently, the approach presented by Zhang et al. achieved much better
184 result, but it involved (a) spatial calibration, (b) detection of dark and bright anatomical
185 structures such as vessels and OD respectively, also (c) bright border regions detection
186 before actual extraction of candidates. Also, there are other techniques based on textu-
187 ral (Morales et al., 2017; Porwal et al., 2018c) and mid-level (Pires et al., 2017) features
188 of retinal images that forgo the lesion segmentation step for DR screening. However,
189 most of these techniques depend on the intermediate steps mentioned above. In the
190 approach based on machine learning (Roychowdhury et al., 2014) detected bright and
191 dark lesions as a first step and later performed the hierarchical lesion classification to
192 generate a severity grade for DR. Similarly, Antal and Hajdu (2014) proposed a strat-
193 egy involving image-level quality assessment, pre-screening followed by lesion and
194 anatomical features extraction to finally decide about the presence of DR using ensem-
195 ble of classifiers. Further, for identification of different stages of DR features from
196 morphological region properties (Yun et al., 2008), texture parameters (Acharya et al.,
197 2012; Mookiah et al., 2013b), non-linear features of the higher-order spectra Acharya
198 et al. (2008), hybrid Dhara et al. (2015) and information fusion (Niemeijer et al., 2009)

199 approaches were found useful. As the DME is graded based on the location of the EXs
200 from macula, many researchers (Giancardo et al., 2012; Medhi and Dandapat, 2014;
201 Perdomo et al., 2016; Marin et al., 2018) proposed EXs based features to determine the
202 severity of the DME. While several others (Deepak and Sivaswamy, 2012; Mookiah
203 et al., 2015; Acharya et al., 2017) have proposed various feature extraction techniques
204 to grade DME stages without segmenting EXs. Mainly for the approaches in this sec-
205 tion, the features are based on the color, brightness, size, shape, edge strength, tex-
206 ture, and contextual information of pixel clusters in spatial and/or transform domain.
207 Whereas the classification is achieved through the classifiers such as K Nearest Neigh-
208 bors (KNN), Naive Bayes, Support Vector Machine (SVM), Artificial Neural Network
209 (ANN), Decision Trees, etc.

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

216 *2.2. Deep learning methods*

217 Deep Learning is a general term to define multi-layered neural networks able to
218 concurrently learn a low-level data representation and higher-level parameters directly
219 from the data. This representation learning capability drastically reduces the need for
220 engineering ad-hoc features, however, the full end-to-end training of deep learning-
221 based approaches typically require a significant number of samples. Its rapid develop-
222 ment in recent times is mostly due to a massive influx of data, advances in computing
223 power and developments in learning algorithms that enabled the construction of multi-
224 layer (more than two) networks (Hinton, 2018; Voulodimos et al., 2018). This progress
225 has induced interests in the creation of analytical, data-driven models based on ma-
226 chine learning in health informatics (Ching et al., 2018; Ravı et al., 2017). Hence, it is
227 emerging as an effective tool for machine learning, promising to reshape the future of
228 automated medical image analysis (Greenspan et al., 2016; Litjens et al., 2017; Suzuki,

229 2017; Shen et al., 2017; Kim et al., 2018; Ker et al., 2018). Among various methodolog-
230 ical variants of deep learning, Convolutional Neural Networks (CNNs or ConvNets) are
231 the most popular within the field of medical image analysis (Hoo-Chang et al., 2016;
232 Carin and Pencina, 2018). Several configurations and variants of CNN's are available
233 in the literature, some of the most popular are AlexNet (Krizhevsky et al., 2012), VGG
234 (Simonyan and Zisserman, 2014), GoogLeNet (Szegedy et al., 2015) and ResNet (He
235 et al., 2016).

236 Deep learning has also been widely utilized in the retinal image analysis because
237 of its unique characteristic of preserving local image relations. Majority of the ap-
238 proaches in the literature employ deep learning to retinal images by utilizing “off-the-
239 shelf CNN” features as complementary information channels to other handcrafted fea-
240 tures or local saliency maps for detection of abnormalities associated with DR (Chudzik
241 et al., 2018; Orlando et al., 2018; Dai et al., 2018), segmentation of OD (Zilly et al.,
242 2017; Fu et al., 2018), and the detection of DR (Rangrej and Sivaswamy, 2017). The
243 authors (Fu et al., 2016) employ fully connected conditional random fields along with
244 CNN to integrate the discriminative vessel probability map and long-range interactions
245 between pixels to obtain final binary vasculature. Whereas some approaches initial-
246 ized the parameters with those of pre-trained models (on non-medical images), then
247 “fine-tuned” (Tajbakhsh et al., 2016) the network parameters for DR screening (Gul-
248 shan et al., 2016; Carson Lam et al., 2018). In another approach researchers used
249 two-dimensional (2D) image patches as an input instead the full-sized images for le-
250 sion detection (Tan et al., 2017b; van Grinsven et al., 2016; Lam et al., 2018; Chudzik
251 et al., 2018; Khojasteh et al., 2018), and OD and fovea detection (Tan et al., 2017a). In
252 (García et al., 2017) trained the “CNN from scratch” and compared it with the fine-
253 tuning results based on the other two existing architectures. Recently, Shah et al.
254 (2018) demonstrated that the ensemble training of auto-encoders stimulates diversity
255 in learning dictionary of visual kernels for detection of abnormalities. Whereas Gian-
256 cardo et al. (2017) proposed a novel way to compute the vasculature embedding that
257 leverages the internal representation of a new encoder-enhanced CNN, demonstrating
258 improvement in the DR classification and retrieval task.

259 There is a significant development in the automated identification of DR using CNN

260 models in recent time. A customized CNN (Gargeya and Leng, 2017) proposed for
261 DR screening and trained using 75,137 obtained from EyePACS system (Cuadros and
262 Bresnick, 2009), where an additional classifier was further employed on the CNN-
263 derived features to determine if the image is with or without retinopathy. Similarly,
264 Google Inc. (Gulshan et al., 2016) developed a network optimized (fine tuning) for im-
265 age classification, in which a CNN is trained by utilizing a retrospective development
266 database consisting of 128,175 images with the labels. There are some hybrid algo-
267 rithms, in which multiple, semi-dependent CNN's are trained based on the appearance
268 of retinal lesions (Abràmoff et al., 2016; Quellec et al., 2016). A step further, the
269 researchers (Quellec et al., 2017) demonstrated an ability of lesion segmentation based
270 on the CNN trained for image level classification. However, Lynch et al. (2017) demon-
271 strated that the hybrid algorithms based on multiple semi-dependent CNNs might offer
272 a more robust option for DR referral screening, stressing the importance of lesion seg-
273 mentation. For further details, readers are recommended to follow recent reviews for
274 detection of exudates (Fraz et al., 2018), red lesions (Biyani and Patre, 2018) and a sys-
275 tematic review with a focus on the computer-aided diagnosis of DR (Mookiah et al.,
276 2013a; Nørgaard and Grauslund, 2018).

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

284 **3. Indian Diabetic Retinopathy Image Dataset**

285 *3.1. Image Acquisition*

286 The IDRiD dataset (Porwal et al., 2018a) was created from real clinical exams ac-
287 quired at an Eye Clinic located in Nanded, (M.S.), India. The fundus photographs of
288 people affected by diabetes were captured with focus on macula using Kowa $VX-10\alpha$

289 fundus camera. Prior to capturing of images, pupils of all subjects were dilated with
290 one drop of tropicamide at 0.5% concentration. The captured images have 50° field of
291 view and resolution of 4288×2848 pixels stored in *jpg* format. The final dataset is
292 composed of 516 images divided into five DR (0 – 4) and three DME (0 – 2) classes
293 with well-defined characteristics according to international standards of clinical rele-
294 vance. It provides expert markups of typical diabetic retinopathy lesions and normal
295 retinal structures. It also provides disease severity level of DR, and DME for each
296 image in the database. Three types of ground-truths are available in the dataset:

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

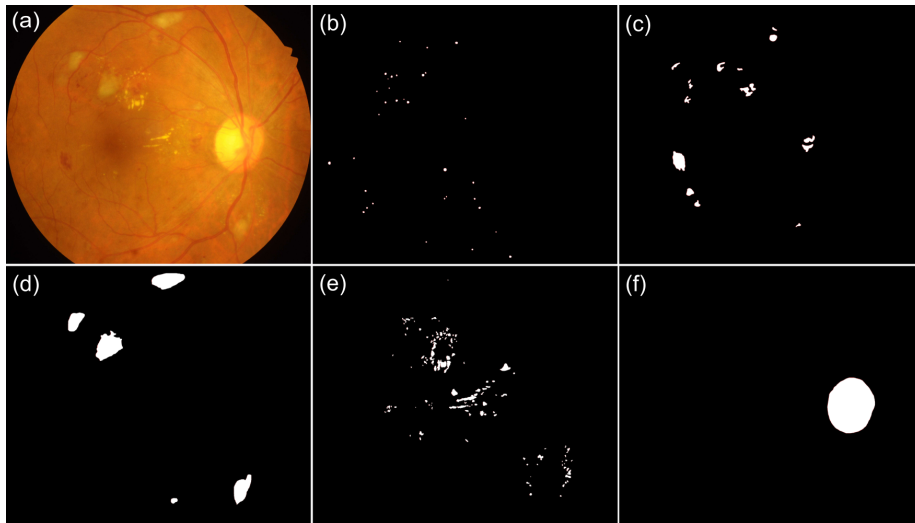


Fig. 2. Retinal photograph and different annotations: (a) sample fundus image from the IDRiD dataset; sample ground truths of (b-f) MAs, HEs, SEs, EXs and OD respectively.

305 *2. Image Level Grading.* It consist of information meant to describe overall risk factor
 306 associated with an entire image. Two medical experts graded the full set of 516 images
 307 with a variety of pathological conditions of DR and DME. Grading for all images is
 308 available in CSV file. The diabetic retinal images were classified into separate groups
 309 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

Table 1. DR Severity Grading.

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: >20 intraretinal HEs Venous beading Intraretinal microvascular abnormalities no signs of PDR
4: PDR	Either or both of the following: Neovascularization Vitreous/pre-retinal HE

310
 311 to macula center region (Decencière et al., 2014) as shown in Table 2.

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

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

314 The IDRiD dataset is available from the IEEE Dataport Repository⁴ under a Cre-
315 ative Commons Attribution 4.0 License. The more detailed information about the data
316 is available in the data descriptor (Porwal et al., 2018b). Tables A.1 and A.2 highlight
317 a comparative strength of the presented dataset with respect to the existing datasets.
318 IDRiD is the only dataset that provides all three types of annotations mentioned above.
319 Streamlining the collection of annotations would allow it to be utilized in research and
320 would lead to better generalizable models for image analysis to be developed, enabling
321 further progress in the automated DR diagnosis.

322 **4. Challenge Organization**

323 The “Diabetic Retinopathy: Segmentation and Grading” challenge was composed
324 into various stages, giving a well-organized work process to potentiate the success of
325 the contest. Fig. 3 depicts the work-flow of the overall challenge organization. The
326 challenge was officially announced at the ISBI - 2018 website⁵ on 15th October 2017.
327 The challenge was subdivided into three sub-challenges as follows:

- 328 1. Lesion Segmentation: Segmentation of retinal lesions associated with DR as
329 MAs, HEs, EXs and SEs.
- 330 2. Disease Grading: Classification of fundus images according to the severity level
331 of DR and DME.
- 332 3. OD detection and Segmentation, and Fovea Detection: Automatic localization
333 of OD and fovea center coordinates, and segmentation of OD.

334 The challenge involved 4 stages, as detailed below:

335 *Stage 1: Data Preparation and Distribution.* The IDRiD dataset was adopted for this
336 challenge, where experts verified that all images are of adequate quality, clinically rele-
337 vant, that no image is duplicated and that a reasonable mixture of disease stratification
338 representative of DR and DME is present. The dataset along with the ground truths
339 were separated into training set and test set. For the images with pixel level annotations,

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

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

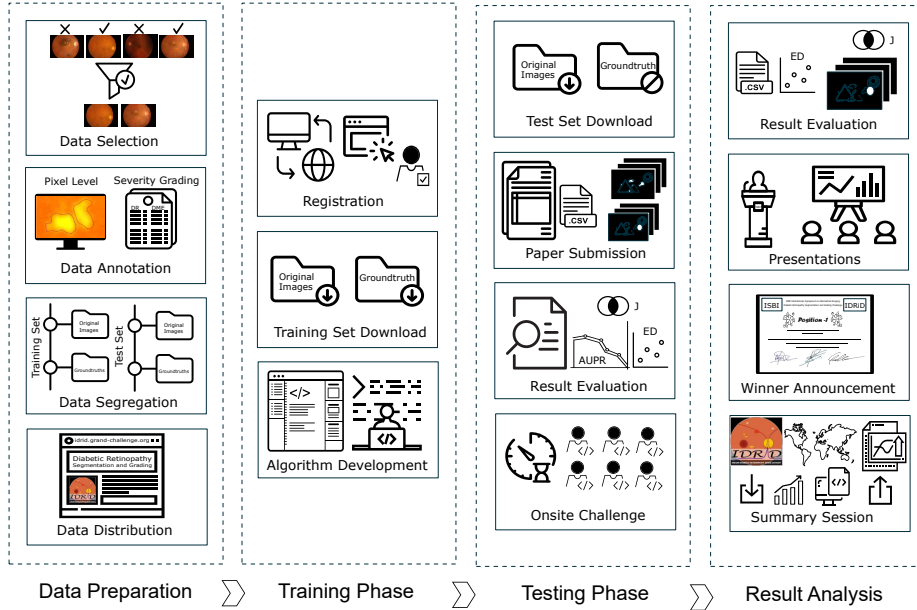


Fig. 3. Workflow of the ISBI - 2018: Diabetic Retinopathy Segmentation and Grading Challenge

340 the data was separated as 2/3 for training (Set-A) and 1/3 for testing (Set-B) (See Table
 341 3). Similarly, data for the OD segmentation (part of sub-challenge – 3) was divided in

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

341
 342 same ratio into Set-A (54 images) and Set-B (27 images). The percentage of images
 343 that should be in each subset for lesion and OD segmentation tasks (sub-challenge – 1
 344 and part of sub-challenge – 3) were chosen based on the research outcome (Dobbin and
 345 Simon, 2011) which demonstrated that splitting data into 2/3 (training): 1/3 (testing)
 346 is an optimal choice for the sample sizes from 50 to 200. For the other sub-challenges
 347 (disease grading, and OD and fovea center locations), data was separated in 80 (train-
 348 ing set: Set-A): 20 (testing set: Set-B) ratio. The percentage of data split in this case is

349 done to provide an adequate amount of data divided into different severity levels. Note
 350 that the dataset was stratified according the DR and DME grades before splitting. A
 351 breakdown of the details of the dataset is shown in Table 4.

Table 4. Stratification of retinal images graded for DR and DME.

DR Grade	Set-A	Set-B	DME 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			

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

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

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

⁶<https://grand-challenge.org/>

369 on-site evaluation of sub-challenge – 1 considering the timing constrains in evaluation
370 of the results for individual image segmentation results.

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

374 1. Sub-challenge – 1: Lesion Segmentation

375 Task - 1: *MA Segmentation*

376 Task - 2: *HE Segmentation*

377 Task - 3: *SE Segmentation*

378 Task - 4: *EX Segmentation*

379 2. Sub-challenge – 2: Disease Grading

380 Task - 5: *DR and DME Grading*

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

382 Task - 6: *OD Center Localization*

383 Task - 7: *Fovea Center Localization*

384 Task - 8: *OD Segmentation*

385 Challenge site was made open for submission from 12th February and participants
386 could submit their results and paper describing their approach till March 11, 2018 to
387 the organizers. Participants could submit up to three methods to be evaluated per team
388 for each task, provided that there was a significant difference between the techniques,
389 beyond a simple change or alteration of parameters. For Tasks 1 to 4 (i.e. sub-challenge
390 – 1) and task-8, the teams were asked to submit output probability maps as grayscale
391 images and for all other tasks it was accepted in CSV format. The submitted results
392 were evaluated by the challenge organizers and their performance was displayed on
393 leaderboard of the challenge website. For sub-challenge – 1, the teams were assessed
394 based on the performance of results submitted on the test set, whereas, for other two
395 sub-challenges assessment was based on the results on the training set obtained through
396 leave one out cross-validation approach. In this phase, it received very good response
397 from the research community with 148 submissions by 37 different teams, out of which
398 16 teams were shortlisted for participation to the on-site challenge. Amongst invited,

399 13 teams confirmed their participation in the on-site challenge, whereas, two teams
400 declined to participate due to other commitments and one team was not able arrange
401 financial support in the limited time.















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

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

425 **5. Competing Solutions**

426 Majority of participating teams proposed a CNN based approach for solving tasks
427 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.

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.	×	×	×	×	×	×	✓	✓	✓

428 to CNN and its variants utilized by the participating teams. Further it summaries the
429 solutions and related technical specifications. For the detailed description of a particu-
430 lar approach please refer to the proceedings of the ISBI Grand Challenge Workshop at
431 https://idrid.grand-challenge.org/Challenge_Proceedings/.

432 For the input image, CNN transforms the raw image pixels on one end to generate a
433 single differentiable score function at the other. It exploits three mechanisms — sparse
434 connections (*a.k.a.* local receptive field), weight sharing and invariant (or equivariant)
435 representation — that makes it computationally efficient (Shen et al., 2017). The CNN
436 architecture typically consists of an input layer followed by sequence of convolutional

437 (CONV), subsampling (POOL), fully-connected (FC) layers and finally a Softmax or
 438 regression layer, to generate the desired output. Functions of all layers are detailed as
 439 follows:

440 The CONV layer comprises of a set of independent filters (or kernels) that are uti-
 441 lized to perform 2D convolution with the input layer (I) to produce the feature (or
 442 activation) maps (A) that give the responses of kernels at every spatial position. Math-
 443 ematically, for the input patch ($I_{x,y}^\ell$) centered at location (x, y) of the ℓ^{th} layer, the
 444 feature value in the i^{th} feature map, $A_{x,y,i}^\ell$, is obtained as:

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

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

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

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

462 The last layer of CNN's is an output layer (O), here the Softmax operator is com-
 463 monly used for the classification tasks. The optimum parameters (θ , common no-
 464 tation for both w and b) for a particular task can be determined by minimizing the

465 loss function (L) defined for the task. Mathematically, for N input-output relations
466 $\{(I^n, O^n); n \in [1, \dots, N]\}$ and corresponding labels G^n the loss can be derived as:

$$L = \frac{1}{N} \sum_{n=1}^N \ln(\theta; G^n, O^n) \quad (2)$$

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

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

478 In this challenge, mainly for disease severity grading problem, participants either
479 directly utilized existing variants of CNN's or ensembled them to demarcate the in-
480 put image to one of the class mentioned above. Several configurations and variants of
481 CNN's are available in literature, some of the most popular are AlexNet (Krizhevsky
482 et al., 2012), VGG (Simonyan and Zisserman, 2014), GoogLeNet (Szegedy et al.,
483 2015) and ResNet (He et al., 2016) due to their superior performance on different
484 benchmarks for object recognition tasks. A typical trend with the evolution of these
485 architectures is that the networks have gotten deeper, e.g., ResNet is about 19, 8 and 7
486 times deeper than AlexNet, VGGNet, and GoogLeNet respectively. While the increas-
487 ing depth improves feature representation and prediction performance, it also increases
488 complexity, making it difficult to optimize and even becomes prone to overfitting. Fur-
489 ther, the increasing number of layers (i.e., network depth) leads to vanishing gradient
490 problems as a result of a large number of multiplication operations. Hence, many
491 teams chose the DenseNet (Iandola et al., 2014) which connects each layer to every




















492 other layer in a feed-forward fashion, reducing the number of training parameters and
493 alleviates the vanishing gradient problem. DenseNet exhibits $\ell(\ell + 1)/2$ connections
494 in ℓ layer network, instead of only ℓ , as in the networks mentioned above. This enables
495 feature reuse throughout the network that leads to more compact internal representa-
496 tions and in turn, enhances its prediction accuracy. Another opted approach, Deep
497 Layer Aggregation (DLA) structures (Yu et al., 2017), extends the “shallow” skip con-
498 nections in DenseNet to incorporate more depth and sharing of the features. DLA uses
499 two structures – iterative deep aggregation (IDA) and hierarchical deep aggregation
500 (HDA) that iteratively and hierarchically fuse the feature hierarchies (i.e. semantic and
501 spatial) to make networks work with better accuracy and fewer parameters. Recent
502 Fully Convolutional Network (FCN) (Long et al., 2015) adapt and extend deep clas-
503 sification architectures (VGG and GoogLeNet) into fully convolutional networks and
504 transfer their learned representations by fine-tuning to the segmentation task. It defines
505 a skip architecture that combines semantic information from a deep, coarse layer with
506 appearance information from a shallow, fine layer to produce accurate and detailed
507 segmentations.

508 For the lesion segmentation task, most of the participating teams exploit U-Net
509 architecture (Ronneberger et al., 2015). The main idea in U-Net architecture is to sup-
510 plement the usual contracting network through a symmetric expansive path by addition
511 of successive layers, where upsampling (via deconvolution) is performed instead of
512 pooling operation. The upsampling part consists of large number of feature channels,
513 that allow the network to propagate context information to higher resolution layers.
514 The high resolution features from the contracting path are merged with the upsampled
515 output and fed to soft-max classifier for pixel-wise classification. This network works
516 with very few training images and enables the seamless segmentation of high resolution
517 images by means of an overlap-tile strategy. Other similar architecture SegNet (Badri-
518 narayanan et al., 2015) was opted by a team, it consists of an encoder and decoder
519 network, where the encoder network is topologically identical to the CONV layers in
520 VGG16 and in which FC layer is replaced by a softmax layer. Whereas, the decoder
521 network comprises a hierarchy of decoders, one corresponding to each encoder. The
522 decoder uses max-pooling indices for upsampling its encoder input to produce a sparse

523 feature maps. Later, it convolves the sparse feature maps with a trainable filter bank to
524 densify them. At last, the decoder output is fed to a soft-max classifier for generation
525 of segmentation map. One team choose Mask R-CNN (He et al., 2017), a technique
526 primarily based on a Region Proposal Network (RPN) that shares convolutional fea-
527 tures of entire image with the detection network, thus enabling region proposals to
528 localize and further segments normal and abnormal structures in the retina. RPN is a
529 fully convolutional network that contributes in concurrently predicting object bounds
530 and “objectness” scores at each position.

531 Following subsections present the solutions designed by participating teams with
532 respect to three sub-challenges. Table 6 summarizes the data augmentation, normaliza-
533 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.

Task	Team Name	Data Augmentation					Other	Data Normalization	Data Preprocessing
		RF	RR	RS	RT	RC			
Sub-challenge - 1	 VRT	✓	✓	✓	✓	✓	shear	✓	FOV cropping, division by 255 then mean subtraction
	 iFLYTEK	✓	✓	✓	✓	✓	×	✓	lesion patch extraction
	 PATech	✓	✓	×	✓	×	color ¹	✓	RGB to LUV, contrast adjustment
	 SDNU	✓	✓	×	×	×	×	-	-
	 SOONER	✓	✓	×	×	✓	×	✓	mean subtraction, lesion patch extraction
	 LzyUNCC	✓	×	×	×	✓	stochastic and photo-metric ²	-	FOV cropping, image enhancement
	 SAIHST	✓	✓	×	×	×	×	✓	CLAHE, Gaussian smoothing
Sub-challenge - 2	 LzyUNCC	✓	×	×	×	✓	color ¹ stochastic and photo-metric ²	-	FOV cropping, image enhancement mean subtraction
	 VRT	×	×	×	×	×	×	✓	mean subtraction
	 Mammoth	✓	✓	✓	✓	×	color	×	morphological opening and closing
	 AVASAVA	✓	×	×	×	✓	×	✓	intensity scaling
	 HarangiM1	×	×	×	×	×	×	✓	FOV cropping
	 HarangiM2	×	×	×	×	×	×	✓	-
Sub-challenge - 3	 DeepDR	×	×	×	×	✓	OD, fovea region	✓	FOV cropping, mean subtraction
	 VRT	✓	✓	✓	✓	✓	shear and cropped OD	✓	FOV cropping, contrast adjustment
	 ZJU-BII-SGEX	×	×	×	×	×	×	✓	FOV cropping
	 SDNU	✓	×	✓	×	×	×	-	-
	 IITkpkLIV	✓	✓	×	×	×	×	✓	-
	 CBER	×	×	×	×	×	×	-	-

¹ Reference: Krizhevsky et al. (2012)

² Reference: Howard (2013)

534 *A. Sub-challenge – 1: Lesion Segmentation*

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

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

542 Son et al. modified U-Net in such a way that the upsampling layers have the same
 543 number of feature maps with the layers concatenated, based on the motivation that fea-
 544 tures in initial layers and upsampled layers are equally important to the segmentation,
 545 thus should have the same number of feature maps. Additionally, they adjusted the
 546 number of max-pooling so that radius of the largest lesion spans a pixel in the most
 547 coarse layer. In case of EX and HE, max-pooling is done 6 times, whereas for SE and
 548 MA it is done 4 times and twice. Further, for dealing with MA’s, they used inverse pixel
 549 shuffling to convert a $1280 \times 1280 \times 3$ pixels image to $640 \times 640 \times 12$ for network input
 550 and pixel shuffling (Shi et al., 2016) to convert $640 \times 640 \times 4$ segmentation map into
 551 $1280 \times 1280 \times 1$ pixels. Later, the pairs of a normalized fundus image and reference
 552 ground truths were fed to the network to generate segmentation result in range $[0,1]$.
 553 They used weighted binary cross entropy (Murphy, 2012) as loss function given by

$$L = \frac{1}{N} \sum_{n=1}^N [-\alpha G^n \log O^n - (1 - G^n) \log(1 - O^n)] \quad (3)$$

554 where N denotes the number of the pairs in a batch, G^n and O^n represent true seg-
 555 mentation and predicted segmentation for n^{th} image. The value of α was determined
 556 as follows:

$$\alpha = \frac{B_0^i}{\gamma F_1^i} \quad (4)$$

557 where B_0^n and F_1^n denote the number of background and foreground pixels in the
 558 n^{th} image. Since background overwhelms foreground in the lesion segmentation, this

559 loss function was designed to penalize false negatives in order to boost sensitivity,
 560 an important factor in detecting lesions. Also, γ was left as a hyper-parameter and
 561 chosen out of $\{0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 256, 512\}$ to yield the highest AU-PR on
 562 the validation set. The final selected γ values for different lesions are summarized in
 Table 7. They trained the network over 300 epochs using Adam optimizer (Kingma

Table 7. γ values in Eq. 4

EXs	SEs	HEs	MAs
64	512	8	32

563
 564 and Ba, 2014) with hyper-parameters of $\beta_1 = 0.5, \beta_2 = 0.999$ and learning rate
 565 of $2e^{-4}$ until 250 epochs and $2e^{-5}$ until the end. All implementation was done
 566 by Keras 2.0.8 with tensorflow backend 1.4.0 using a server with 8 TITAN X (pas-
 567 cal). The source code is available at [https://bitbucket.org/woalsdnd/](https://bitbucket.org/woalsdnd/isbi-2018-fundus-challenge)
 568 [isbi-2018-fundus-challenge](https://bitbucket.org/woalsdnd/isbi-2018-fundus-challenge).

569 A.2. *iFLYTEK-MIG* (Fengyan Wang et al.)

570 Wang et al. proposed a novel cascaded CNN based approach for retinal lesion
 571 segmentation with U-Net as a base model. It consists of three stages, the first stage
 572 is a coarse segmentation model to get initial segmentation masks, then second stage
 573 is a cascade classifier which was designed for false positive reduction, at last a fine
 574 segmentation model was used to refine results from the previous stages. First stage
 575 model was trained using the patches of size 256×256 pixels centered on the particular
 576 lesion amongst MA, HE or EX and 320×320 pixels for SE, resulting in the coarse
 577 segmentation outcome. Results of previous stage are coarse due to the fact that non-
 578 focus regions (non target lesions) were not utilized in the learning process leading to
 579 high false positive count. In the second stage, unlike the first segmentation model
 580 which used a lesion centered sample from input dataset pool, candidate regions were
 581 extracted using probability maps from the previous stage. Here, the input size fed to
 582 model for SE was $320 \times 320 \times 3$ pixels, for HE and EX it was $256 \times 256 \times 3$ pixels,
 583 and for MA it was modified to $80 \times 80 \times 3$ pixels considering its small appearance. In
 584 this step, a candidate region was regarded as a positive sample if its intersection-over-

585 union with the ground truth was greater than the given threshold (i.e. 0.5). In this way,
 586 most trivial non-focus regions were effectively rejected. However, it was identified
 587 in the test that a small proportion of false positives still exist, so an additional model
 588 was introduced to refine the segmentation results. In the last stage, candidate regions
 589 survived from the second stage were utilized as the input patches resulting in more
 590 accurate segmentation results. For first and third stage, they used binary cross entropy
 591 or dice loss function (multi-model training), whereas, for second stage, they used only
 592 binary cross entropy as loss function. The first, second and third stage models were
 593 trained for 100, 300 and 100 epochs respectively with momentum of 0.9. In which,
 594 the initial learning rate for first and third stage was set 0.1 and is reduced by 10 times
 595 every 30 epochs, and for second stage it was set to 0.001 reduced by 10 times every 80
 596 epochs. MXNET platform was used for training the models.

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

598 Lihong et al. developed a novel patch-based CNN model (as shown in Fig. 4) in
 599 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

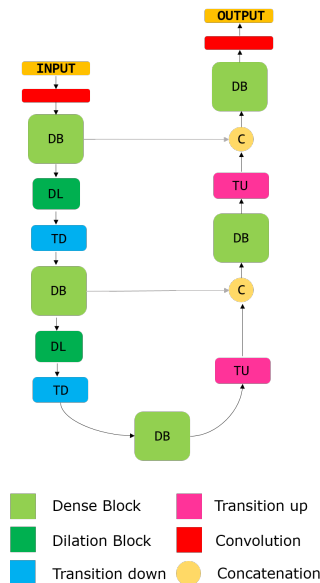


Fig. 4. Proposed architecture for lesion segmentation

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

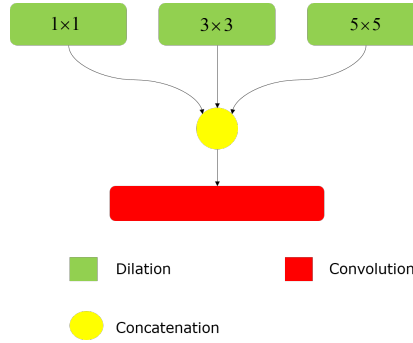


Fig. 5. Architecture for dilation block.

608 Initially, they extracted regions within FOV from the images and then normalized
 609 them to eliminate local contrast differences and uneven illumination. Later, they used
 610 small patches 256×256 pixels at stride of 64 (128 for MA) to generate the training
 611 samples (only patches that overlap with the lesion ground truth) followed by data aug-
 612 mentation before feeding to the model. To deal with highly imbalanced spread of data,
 613 they designed a loss function that is combination of dice function (Sudre et al., 2017)
 614 and 2D cross Entropy as follows:

$$\begin{aligned}
 L = & -mean(w_{10} * G * \log(O) \\
 & + w_{11} * (1 - G) * \log(1 - O) \\
 & + w_2 * dice(G))
 \end{aligned} \tag{5}$$

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

617 and cross entropy loss. The values of w_{10} , w_{11} and w_2 were empirically set to 0.7,
618 0.3 and 0.4 respectively. The models were trained using Adam optimizer with default
619 parameters, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate was set to 2×10^{-4} , and
620 then divided by 20 in every 20 epochs. This model was implemented with pytorch1.12
621 and Tesla M60 platform was utilized for training on the centos 7.2 operating system.

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

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

644 *A.5. LZYUNCC (Zhongyu Li et. al.)*

645 Li et al. developed method based on FCN by embedding DLA structure for the seg-
646 mentation of HE’s and SE’s. As the lesions are located dispersively and irregularly, em-
647 bedding of DLA structure with FCN enables better aggregation of semantic and spatial
648 information from local and global level provides a boost in recognizing their presence.
649 They used retinal images with pixel-level ground truth annotations from both IDRiD
650 and E-Ophtha database. They first adopted a series of methods for data preprocessing
651 and augmentation. Subsequently, considering the correlation between EX’s and SE’s,
652 they first trained an initial model for the segmentation of EX. They chose a smaller
653 model, i.e., DLA-34 to train the segmentation network with binary cross entropy as
654 a loss function. At last, the trained deep model was fine-tuned for the segmentation
655 of SE. While the model training of EX segmentation, a trade-off parameter (penalty)
656 was assigned in the loss function to control the weights of foreground pixels, and tried
657 different penalty value from 1 to 16 during the model training. At last, these segmen-
658 tation results were fused to adaptively compute the best performance. They adopted
659 the original DLA cityscapes segmentation experimental settings (Yu et al., 2017) and
660 trained the model for 100 epochs with batch size 4, where the poly learning rate was
661 $(1 - \frac{epoch-1}{totalepoch})^{0.9}$ with momentum of 0.9. The initial learning rate was set to 0.01.

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

663 Choi et al. proposed a model for segmentation of EX based on U-net, in which
664 the convolution layers of the encoder path are replaced with dense blocks. Whereas,
665 the decoder path of their model was kept identical to that of general U-net. They
666 built the dense block with growth factor of 12 and 3×3 convolution layers, batch
667 normalization, and ReLU activation. The last layer generates pixel level prediction map
668 for EXs through the sigmoid activation function. For training, they utilized only green
669 channel of fundus image and enhanced it using Contrast Limited Adaptive Histogram
670 Equalization (CLAHE). Later, each image was padded to a size of 4352×3072 pixels
671 and cropped into 204 patches of 512×512 pixels. These patches are further augmented
672 and used for training. The losses were calculated by the binary cross-entropy. The
673 model was trained for 20 epochs with a mini-batch size of 10 and they used Adam

674 optimizer with an initial learning rate of $2e^{-4}$, β_1 of 0.9 and β_2 of 0.999. The model
675 was programmed in Keras 2.1.4 served with Tensorflow 1.3.0 backend.

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

677 Sui et al. proposed a method based on Mask R-CNN structure to segment lesions
678 from the fundus image. They adopted implementation of Mask R-CNN from (Ab-
679 dulla, 2017) for solving the problem. This method could detect different objects while
680 simultaneously generating instance segmentation mask.

681 Network training precedes the data augmentation process and binary cross entropy
682 was used as a loss function. The initial learning rate was set to 0.02 with momentum of
683 0.9. They chose ResNet-101 as a backbone. They implemented algorithm in Keras with
684 Tensorflow as backend and processed on 8 NVIDIA TITAN Xp GPUs. The experiment
685 environment was built under Ubuntu 16.06.

686 B. Sub-challenge – 2: Disease Grading

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

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

691 Li et al. developed method based on the ResNet by embedding DLA structure for
692 the automated grading of DR and DME. For this work they used IDRiD and Kaggle
693 dataset. Initially, for the given training images, they perform data preprocessing and
694 data augmentation. Subsequently, based on the designed ResNet with DLA structure,
695 initial models are trained using 35,000 retinal images from the Kaggle dataset. Later,
696 they fine-tuned the model using the IDRiD dataset through 5 fold cross validation tech-
697 nique. Finally, the 5 outputs are ensembled together as the final grades for input im-
698 ages. It is important to note that networks for the grading of DR and DME were trained
699 separately. The training was performed by Stochastic Gradient Descent (SGD) with a
700 mini-batch size of 64, while the learning rate starts from 0.001 and is then divided by
701 10 every 20 epochs, for 30 epochs in total. The other hyper-parameters are fixed to the
702 settings of original DLA ImageNet classification (Yu et al., 2017).

703 *B.2. VRT (Jaemin Son et al.)*

704 Son et al. used network (Son et al., 2018) for DR grading. Kaggle dataset was ini-
705 tially used to pre-train the network and then the model was fine-tuned with the IDRiD
706 data. Penultimate layer was Global Average Pooled (GAP) and connected with FC
707 layer. The entire output is a single value from which L2 loss was calculated against
708 the true label. SGD was used with nesterov momentum of 0.9 as optimizer. Learn-
709 ing rate was set to 10^{-3} . The model was trained for 100 epochs. Fundus image was
710 normalized in range $[0, 1]$ and the mean was subtracted channel-wise. For grading of
711 DME, the segmented EXs (using the segmentation network proposed in sub-challenge
712 – 1), localized fovea and segmented OD (using the segmentation network proposed in
713 sub-challenge – 3) were utilized for making final decision. With these information,
714 semi-major axis of the segmented OD (r) was estimated. Further, the fundus image
715 is divided into three regions as macular region: $\|x - c\| < r$, near macular region:
716 $r < \|x - c\| < 2r$ and remaining region: $2r < \|x - c\|$. where x denotes a point in the
717 image.

718 Furthermore, several features such as sum of intensity for segmented EX, the num-
719 ber of pixels above the threshold (178 in the $[0, 255]$ scale), the number of pixels
720 for the smallest and largest blob, the mean pixel numbers of blobs are extracted for
721 each area, and binary flag that indicates whether the OD is segmented. Now, features
722 with high importance were selected among numerous features in the initial training
723 due to gradient boosting (for instance, XGBoost) was likely to overfit when provided
724 with overly redundant features. Messidor dataset was added to the given data and
725 out of which 10% of images were left as validation set. Sets of hyper-parameters are
726 searched by grid-search. The combination of hyper parameters that yielded the highest
727 accuracy in validation set was min child eight: 2, subsample: 0.2, colsample by tree:
728 0.2, λ : 9.0, α : 1.0, and depth: 6. Other hyper-parameters are set to default values. All
729 implementations were done by pytorch v0.4.1 using a server with 8 TITAN X (pas-
730 cal). The source code is available at [https://bitbucket.org/woalsdnd/
731 isbi-2018-fundus-challenge](https://bitbucket.org/woalsdnd/isbi-2018-fundus-challenge).

732 *B.3. Mammoth (Junyan Wu et al.)*

733 Wu et al. proposed an unified framework that combines deep feature extractor and
734 statistical feature blending to automatically predict the DR and DME severity scores.
735 For DME, they used DenseNet to directly predict severity score. Whereas for DR,
736 Kaggle training dataset was used to pre-train the DenseNet model through a dynamic
737 sampling mechanism to balance the training instances and later fine tuned using the
738 IDRiD dataset. Initially, the background of all images was cropped and resized to
739 512×512 pixels. Later, morphological opening and closing are utilized to preserve
740 bright and dark regions. For instance, the morphological opening can erase the EXs and
741 highlight the MAs. Whereas, the closing operation can remove MAs and preserve EXs.
742 These operations can be used to denoise specific levels of classifications, for example,
743 the risk of DME only depends on the location of the EXs. Further, several standard
744 data augmentation methods (as shown in Table 6) are also employed. Mean Squared
745 Error (MSE) and cross-entropy with five classes were the loss functions employed to
746 train the network and SGD for optimization. The initial learning rate was set to 0.0005
747 with decrement of 0.1 after every 30 epochs. The initial training was done by 200
748 epochs and fine tuning by 50 epochs. Afterwards, the last layer was removed before
749 final prediction, and its statistical features were aggregated together into a boosting
750 tree. Specifically, 50 pseudo random augmentations were performed to get 50 outputs
751 from last second FC layer (size of 4096), then the mean and standard deviation of 50
752 feature vectors for each image were computed, and both vectors were then concatenated
753 together for training in LightGBM. The output from second last layer of fine-tuning
754 experiments were used to train a blending model, strategy adopted from team o.O's
755 solution of Kaggle DR challenge. Finally, for the disease grading prediction, gradient
756 boosting tree model was built on combined second last layer from pre-trained network
757 and fine-tuned network.

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

759 Harangi et al. proposed an approach for the classification of retinal images via
760 the fusion of two AlexNet, and GoogLeNet. For this aim, they removed a FC and
761 classification layers and interconnect them by inserting a joint FC layer followed by the

762 classic softmax/ classification layers for the final prediction. In this way, single network
763 architecture was created which allows to train the member CNNs simultaneously. For
764 each $I^{(n)}$, let us denote the outputs of the final FC layers of the member CNNs by
765 $\hat{O}_1^{(n)}, \hat{O}_2^{(n)}$. The FC layer of their ensemble aggregates them via

$$\acute{O}^{(n)} = A_1 \hat{O}_1^{(n)} + A_2 \hat{O}_2^{(n)} \quad (6)$$

766 where the weight matrices A_1, A_2 were of size 5×5 and initialized as

$$A_1 = A_2 = \begin{bmatrix} 1/5 & 0 & 0 & 0 & 0 \\ 0 & 1/5 & 0 & 0 & 0 \\ 0 & 0 & 1/5 & 0 & 0 \\ 0 & 0 & 0 & 1/5 & 0 \\ 0 & 0 & 0 & 0 & 1/5 \end{bmatrix} \quad (7)$$

767 The last two layers of the ensemble were a softmax and a classification one. Let
768 $O_{SM}^{(n)}$ be the output of the former layer, the MSE was used for optimization as a loss
769 function:

$$MSE = \frac{1}{2N} \sum_{n=1}^N (\acute{O}_{SM}^{(n)} - O^{(n)})^2 \quad (8)$$

770 During the training phase, back-propagation is applied to minimize the loss via
771 adjusting all the parameters of the member CNNs and the weight matrices A_1, A_2 .

772 For the grading of DME, the final layers of the member CNNs consist of 3 neurons,
773 and the weight matrices A_1, A_2 were 3×3 , initialized as

$$A_1 = A_2 = \begin{bmatrix} 1/3 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 1/3 \end{bmatrix} \quad (9)$$

774 For training they merged the IDRiD and Kaggle training set. The parameters of
775 the architectures were found by the SGD algorithm in 189 and 50 epochs respectively
776 for the DR and DME classification tasks. Learning rate was set to 0.0001. Training

777 times required on the datasets for DR and DME were 96.6 (189 epochs) and 23.4 (50
778 epochs) hours respectively. Implementation of this work was done in Matlab 2017b.
779 Training was performed using an NVIDIA TITAN X GPU card with 7 TFlops of single
780 precision performance, 336.5 GB/s of memory bandwidth, 3,072 CUDA cores, and 12
781 GB memory.

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

783 Alex et al. used ensembles of pre-trained CNNs (on ImageNet dataset), namely,
784 ResNets and DenseNets for the task of disease grading. For the task of grading of DR,
785 two ensembles of CNNs namely “primary” and “expert” classifiers were used. The
786 primary classifier was trained to classify a fundus image as one of the 4 classes viz;
787 Normal, Mild NPDR, Moderate NPDR or S-(N)-PDR, a class formed by clubbing Se-
788 vere NPDR and PDR. The expert classifier was trained exclusively on Severe NPDR
789 or PDR images and was utilized to demarcate the input image as one of the aforemen-
790 tioned classes. During inference, each fundus image was resized to a dimension of
791 256×256 pixels. For the task of grading of DR in fundus images, they used test time
792 augmentation through the “Ten Crop” function defined in PyTorch. The images were
793 first passed through the primary classifier and then through the expert classifier, only if
794 the image was classified as S-(N)-PDR by primary classifier. The final prediction was
795 achieved by using a majority voting scheme.

796 For DME grading, two ensembles were trained in a one versus rest approach. En-
797 semble 1 was trained to classify the input as either “image with no apparent EXs”
798 (Grade 0) or “presence of EXs in image” (Grade 1 & Grade 2), while the Ensemble
799 2 was trained to classify an image as “Grade 2” DME or not (Grade 0 & Grade 1).
800 During inference, the resized images were fed to both ensembles and the final predic-
801 tion was obtained by combining the two predictions by utilizing a set of user defined
802 rules. Briefly, the user defined rules were: an image was classified as Grade 0 DME
803 if ensemble 1 and ensemble 2 predict the absence of EXs and the absence of grade 2
804 DME respectively. A scenario wherein ensemble 2 predicts the presence of grade 2
805 DME, the images were classified under the category “Grade 2 DME” irrespective of
806 the prediction from ensemble 1. Lastly, images were classified as Grade 1 DME if none

807 of the above conditions were satisfied.

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

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

817 Harangi et al. combined self-extracted, CNN-based features with traditional, hand-
818 crafted ones for disease classification. They modified AlexNet to allow the embedding
819 of handcrafted features via a FC layer. In this way, they created a network architec-
820 ture that could be trained in the usual way and additionally uses domain knowledge.
821 They extended the FC layer FC_{fuse} originally containing 4096 neurons of AlexNet by
822 adding 68-dimensional vector containing handcrafted features. Then, the 4164×5 (or
823 4164×3 for DME) layer FC_{class} was considered for the DR (or DME) classification
824 task. In this way, both the final weighing FC_{class} of the handcrafted features were
825 obtained and the 4096 AlexNet features were trained by back propagation.

826 To obtain the 68 handcrafted features used by the CNN, they employed one image
827 level and two lesion specific methods. The amplitude-frequency modulation (AM-
828 FM) method extracts information from an image by decomposing its green channel at
829 different scales into AM-FM components (Havlicek, 1996). As a result, a 30-element
830 feature vector was obtained, which reflects the intensity, geometry and texture of the
831 structures contained in the image (Agurto et al., 2010). Whereas to extract features
832 related to the lesions MA and EX, they employed two detector ensembles (Antal and
833 Hajdu, 2012; Nagy et al., 2011), which consist of a set of <preprocessing method (PP),
834 candidate extractor (CE)> pairs organized into a voting system. Such a <PP, CE> pair
835 was formed by applying the PP to the retinal image and the CE to its output. This way,
836 a <PP, CE> pair extracts a set of lesion candidates from the input image, acting like

837 a single detector algorithm. They used output of these ensembles to obtain 38 features
838 related to the number and size of MA's and EX's. The parameters of the architectures
839 were optimized by SGD algorithm in 85 and 50 epochs for DR and DME respectively.
840 Training times were 83.1 (85 epochs) and 46.2 (50 epochs) hours on the datasets for DR
841 and DME. Implementation of this work was done in Matlab 2017b. Training has been
842 performed using an NVIDIA TITAN X GPU card with 7 TFlops of single precision,
843 336.5 GB/s of memory bandwidth, 3,072 CUDA cores, and 12 GB memory.

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

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

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

849 Dai et al. proposed a novel deep localization method, which allows coarse-to-fine
850 feature encoding strategy for capturing the global and local structures in fundus images,
851 to simultaneously model the two-task learning problem of the OD and fovea localiza-
852 tion. They took advantage of the prior knowledge such as the number of landmarks and
853 their geometric relationship to reliably detect the OD and fovea. Specifically, they first
854 designed a global CNN encoder (with a backbone network of ResNet-50) to localize
855 the OD and fovea centers as a whole by solving a regression task. All max pooling
856 layers were replaced with average pooling layers as compared to the original ResNet
857 architecture, due to the fact that the max pooling could lose some useful pixel-level
858 information for the regression to predict the coordinates. This step was used to si-
859 multaneously perform the two detection tasks, because of the geometric relationship
860 between OD and fovea, the performance of multi-task learning is better than single
861 task. The predicted output coordinates of this global CNN encoder component were
862 used for detecting the bounding boxes of the target OD and fovea. Then the current
863 center coordinates are refined through a local encoder (with a backbone network of
864 VGG-16) which only localizes the OD center or fovea center of their related bounding
865 boxes. During training stage, they designed the effective data augmentation scheme to

866 solve the problem of insufficient training data. In particular, to build the training set
867 of the local encoder, the bounding boxes were randomly selected based on the ground
868 truth, for each object several bounding boxes of different positions and scales were
869 cropped. The local encoder can be reused multiple times to approximate the target co-
870 ordinates. The local encoder was iterated twice for refining centers comprehensively.
871 All three models were initialized from the pre-trained ImageNet network, and replaced
872 the network’s last FC layer and softmax layer by the center coordinates regressor. The
873 regression loss for the center location was the Euclidean loss. The modified loss func-
874 tion for global and local encoders was $0.045(L_{OD} + L_{fovea})$ and $0.045(L_{OD}/L_{fovea})$
875 respectively. Where L_{OD} and L_{fovea} are losses for OD and fovea, and scaling factor
876 was introduced since the original Euclidean distance is too large in practice to con-
877 verge. The proposed learning model was implemented in Caffe framework and trained
878 using SGD with momentum. The FC layers for center regression were initialized from
879 zero-mean Gaussian distributions with standard deviations 0.01 and 0.001. Biases were
880 initialized to 0. The global encoder was trained for 200 epochs, local encoders (OD and
881 fovea both) for 30 epochs respectively. The batch size for the global encoder was 16,
882 and 64 for the other two local encoders. The learning rate was set as 0.01 and was
883 divided by 10 when the error plateaus.

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

885 Son et al. proposed an OD segmentation model consisting of U-Net and CNN
886 that takes a vessel image and outputs 20×20 activation map whose penultimate layer
887 is concatenated to bottleneck layer of the U-Net. Initially, the original images were
888 cropped (3500×2848 pixels), padded (3500×3500 pixels) and then resized ($640 \times$
889 640 pixels). Each image was standardized with its mean and standard deviation (std).
890 When calculating the mean and std, values less than 10 (usually artifacts in the black
891 background) are ignored. Vessel images were prepared with an external network Son
892 et al. (2017). Pixel values in a vessel image range from 0 to 1. It uses external datasets
893 DRIONS-DB (Carmona et al., 2008) and DRIVE (Staal et al., 2004) available with
894 OD and vessel ground truths respectively. For augmentation, the fundus images were
895 affine-transformed and additionally OD was cropped and randomly placed on the image

896 for random number of times (0 to 5). This augmentation was done to prevent the
897 network from segmenting OD solely by brightness. Pairs of a fundus image and the
898 vessel segmentation were provided as input and OD segmentations in the resolution of
899 640×640 and 20×20 pixels are given as the ground truth. Binary cross entropy is used
900 as loss function for both U-Net and vessel network with the loss of $L_{total} = L_{U-Net} +$
901 $0.1 * L_{vessel}$. Total 800 epochs are trained via Adam optimizer and decreasing learning
902 rate with hyper-parameters of $\beta_1 = 0.5, \beta_2 = 0.999$. The learning rate was $2e^{-4}$ until
903 400 epochs and $2e^{-5}$ until the end. Weights and biases were initialized with Glorot
904 initialization method (Glorot and Bengio, 2010).

905 They also proposed a four branch model in which two branches were dedicated
906 to prediction of locations for OD and fovea from vessels (vessel branches) and
907 other two branches aim to predict the locations from both fundus and vessels (main
908 branches). Similar to OD segmentation, penultimate layers of vessel branches were
909 depth-concatenated to the main branches. After deriving an activation map that repre-
910 sents probability of containing the anatomical landmark, hard-coded matrix was mul-
911 tiplied to yield co-ordinates. Original images were cropped as in the segmentation
912 task and standardized with the identical method and later augmented by flip and rota-
913 tion to ease the implementation efforts. Mean absolute error was used as loss function
914 for both outputs with the loss of $L_{total} = L_{main} + 0.3 * L_{vessel}$. SGD was used
915 with nestrov momentum of 0.9 as optimizer. Learning rate was set to 10^{-3} from 1st
916 to 500th epochs and 10^{-4} from 501th to 1000th epochs. All implementation were
917 done in Keras 2.0.8 with tensorflow backend 1.4.0 using a server with 8 TITAN X
918 (pascal). Source code is available at [https://bitbucket.org/woalsdnd/](https://bitbucket.org/woalsdnd/isbi-2018-fundus-challenge)
919 [isbi-2018-fundus-challenge](https://bitbucket.org/woalsdnd/isbi-2018-fundus-challenge).

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

921 Lyu et al. utilized Mask R-CNN to localize and segment OD and fovea simultane-
922 ously. It scans the image and generates region proposals by 2D bounding boxes. Then
923 the proposals were classified into different classes and compute a binary mask for each
924 object. They firstly preprocessed the original retinal image into fixed dimensions as
925 network input. A feature extractor (ResNet-50) with feature pyramid networks (FPN)

926 generates feature maps at different scales, which could be used for regions of interest
927 (ROI) extraction. Then a region proposal network (RPN) scans over the feature maps
928 and locates regions that contain objects. Finally, a ROI head network (RHN) is em-
929 ployed to obtain the label, mask, and refined bounding box for each ROI. They also
930 incorporated prior knowledge of retinal image as a post-processing step to improve the
931 model performance. They used IDRiD dataset and two subsets in RIGA dataset (Al-
932 mazroa et al., 2018) (Messidor and BinRushed, 605 images) with OD mask provided.
933 They applied transfer learning technique to train the model. They firstly trained the
934 RHN network by freezing all the layers of FPN and RPN networks and then fine-tuned
935 all layers. The model was implemented on Tensorflow 1.3 and python 3.4 (source code
936 was modified from Abdulla (2017)). The learning rate started from 0.001 and a mo-
937 mentum of 0.9 was used. The network was trained on one GPU (Tesla K80) with 20
938 epochs.

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

940 Saha et al. used SegNet for segmentation of lesions and OD. OD was added as an
941 additional class in the same problem as lesion segmentation, so that the model could
942 better differentiate EXs and OD which have similar brightness levels. However, in
943 contrast to original SegNet, the final decoder output is fed to a sigmoid layer to produce
944 class probabilities for each pixel independently in 7 channels. Each channel has the
945 same size as input image : 536×356 pixels and consists of activations in the range $[0,1]$
946 where 0 corresponds to background and 1 to the presence of corresponding class. Apart
947 from 5 classes i.e. MA, HE, SE, EX and OD, two additional classes: (i) retinal disk
948 excluding the lesions and OD, and (ii) black background form the 7 channels. Images
949 were downsampled to 536×356 pixels, preserving the aspect ratio. Additionally,
950 Drishti-GS (Sivaswamy et al., 2014) dataset was used for data augmentation to account
951 for case of absence of lesions. Further, horizontal, vertical and 180 degree flipped
952 versions of the original images were taken. The network was trained using binary cross
953 entropy loss function and Adam optimizer with learning rate 10^{-3} and $\beta = 0.9$. Early
954 stopping of the training based on the validation loss is adopted to prevent overfitting.
955 It was observed that the validation loss started to increase after 200 epochs. One more

956 softmax layer is introduced after the Sigmoid layer for normalizing the value of a pixel
957 for each class across channels. Segmented output is finally upsampled for each class to
958 4288×2848 pixels. All implementations were done in PyTorch using 2x Intel Xeon
959 E5 2620 v3 processor with GTX TitanX GPU 12 GB RAM and 64 GB System RAM.

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

961 Sui et al. used Mask R-CNN for solving all tasks in this sub-challenge. Mask R-
962 CNN could realize accurate target detection based on proposed candidate object bound-
963 ing boxes of a RPN to achieve the objective of OD and Fovea localization. At the same
964 time, it could also get the OD segment at the mask predicting branch. The head archi-
965 tecture of Mask R-CNN (ResNet-101 as a backbone) consists of three parallel branches
966 for classification, bounding-box regression, and predicting mask. By this method, the
967 localization of OD and fovea, and segment the mask of OD could be obtained directly.
968 They retrained the network to get the new weight parameter of the framework. During
969 the training phase, the dataset of this challenge was augmented by flipping, resizing
970 and trained by 10-fold cross-validation. After training 2000 epochs, the last trained
971 model is obtained. They implemented this algorithm in Tensorflow and it is processed
972 on 8 NVIDIA TITAN Xp GPUs. The experiment environment is built under Ubuntu
973 16.06.

974 C.6. CBER (Ana Mendonça et al.)

975 Mendonça et al. proposed handcrafted features based approach for the localization
976 and segmentation tasks in this sub-challenge. Distinct methodologies have been devel-
977 oped for detecting and segmenting these structures, mainly based on color and vascular
978 information. The methodology proposed in the context of this challenge includes three
979 inter-dependent modules. Each module performs a single task: OD localization, OD
980 segmentation or fovea localization. While the modules responsible for the OD localiza-
981 tion and segmentation were an improved version of two methods previously published
982 (Mendonca et al., 2013; Dashtbozorg et al., 2015), the method proposed for fovea local-
983 ization was completely new. Initially, the module associated with the OD localization
984 receives a fundus image and segments the retinal vasculature. Afterwards, the entropy

985 of the vessel directions is computed and combined with the image intensities in order
 986 to find the OD center coordinates. For OD segmentation, the module responsible for
 987 this task uses the position of the OD center for defining the region where the sliding
 988 band filter (Pereira et al., 2007; Esteves et al., 2012) is applied. The positions of the
 989 support points which give rise to the maximum filter response were found and used
 990 for delineating the OD boundary. Since a relation between the fovea-OD distance and
 991 the OD diameter was known (Jonas et al., 2015), the module responsible for the fovea
 992 localization begins by defining a search region from the OD position and diameter. The
 993 fovea center is then assigned to the darkest point inside that region.

994 6. Evaluation Measures

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

997 A. Sub-challenge – 1

998 This sub-challenge evaluates the performance of the algorithms for different lesion
 999 segmentation tasks, from the submitted grayscale images, using the available binary
 1000 masks. As in the lesion segmentation task(s) background overwhelms foreground, a
 1001 highly imbalanced scenario, the performance of this task was measured using area
 1002 under precision (*a.k.a.* Positive Predictive Value (PPV)) recall (*a.k.a.* Sensitivity (SN))
 1003 curve (AUPR) (Saito and Rehmsmeier, 2015).

$$SN = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (10)$$

$$PPV = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (11)$$

1004 The curve was obtained by thresholding the results at 33 equally spaced instances
 1005 i.e. $[0, 8, 16, \dots, 256]$ in gray levels or $[0, 0.03125, 0.0625 \dots, 1]$ in probabilities. The
 1006 AUPR provides a single-figure measure (*a.k.a.* mean average precision (mAP)), com-
 1007 puted over the Set-B, was used to rank the participating methods. This performance

1008 metric was used for object detection in The PASCAL Visual Object Classes (VOC)
 1009 Challenge (Everingham et al., 2010). The AUPR measure is more realistic (Boyd et al.,
 1010 2013; Saito and Rehmsmeier, 2015) for the lesion segmentation performance over the
 1011 Area under Receiver Operating Characteristics (ROC).

1012 B. Sub-challenge – 2

1013 Let the expert labels for DR and DME be represented by $DR_G(n)$ and $DME_G(n)$.
 1014 Whereas, $DR_O(n)$ and $DME_O(n)$ are the predicted results, then correct instance is
 1015 the case when the expert label for DR and DME matches with the predicted outcomes
 1016 for both DR and DME. This was done since, even with presence of some exudation that
 1017 may be categorized as mild DR, its location on the retina is also important governing
 1018 factor (to check DME) to decide overall grade of disease. For instance, EXs presence in
 1019 the macular region can affect vision of the patient to greater extent and hence, it should
 1020 be dealt with priority for referral (that may otherwise be missed or cause delay in
 1021 treatment with the present convention of only DR grading) in the automated screening
 1022 systems. Hence, disease grading performance accuracy for this sub-challenge, from
 1023 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  $n = 1, 2, \dots, N$  do
2  |   Correct = 0;
3  |   if ( $DR_O(n) == DR_G(n)$ ) and ( $DME_O(n) == DME_G(n)$ ) then
4  |   |   Correct = Correct + 1;
5  |   end
6  end
7  Average Accuracy =  $\frac{\text{Correct}}{N}$ 

```

1024

1025 C. Sub-challenge – 3

1026 For the given retinal image, the objective of sub-challenge – 3 (task - 6 and 7) was
 1027 to predict the OD and fovea center co-ordinates. The performance of results submitted

1028 in CSV format was evaluated by computing the Euclidean distance (ED) (in pixels)
1029 between manual (ground truth) and automatically predicted center location. Lower ED
1030 indicates better localization. After determining Euclidean distance for each image in
1031 the Set-B, i.e. for 103 images, the average distance representing the whole dataset was
1032 computed and used to rank the participating methods.

1033 The optic disc segmentation (task - 8) performance is evaluated using Jaccard in-
1034 dex (J) (Jaccard, 1908). It represents the proportion of overlapping area between the
1035 segmented OD (O) and the ground truth (G).

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

1036 Higher J indicates better segmentation. For the segmented results, images in range
1037 $[0, 255]$, it was computed at 10 different equally spaced thresholds $[0, 0.1, \dots, 0.9]$
1038 and averaged to obtain final score.

















1039 7. Results

1040 This section reports and discusses the results of all sub-challenges. Performance
1041 of all competing solutions on the Set-B for all eight subtasks are divided into three
1042 sub-challenge categories and discussed including their leaderboard rank.

1043 A. Sub-challenge – 1

1044 In this section, we present the performance of all competing solutions for the lesion
1045 segmentation task. All results received from the participating teams were analyzed
1046 using the validation measure given in section 6.A. This measure generated a set of
1047 precision-recall curves for each of the different techniques. A total of 22 solutions were
1048 evaluated for this sub-challenge (a complete list is available on the challenge website)
1049 and ranked using the area under precision-recall curve values. Amongst them, only
1050 top-4 teams per lesion segmentation task were invited for the challenge workshop and
1051 top-3 teams having overall better performance, the solutions developed by the teams
1052 that ranked amongst top three for at least three different lesion segmentation tasks,
1053 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).

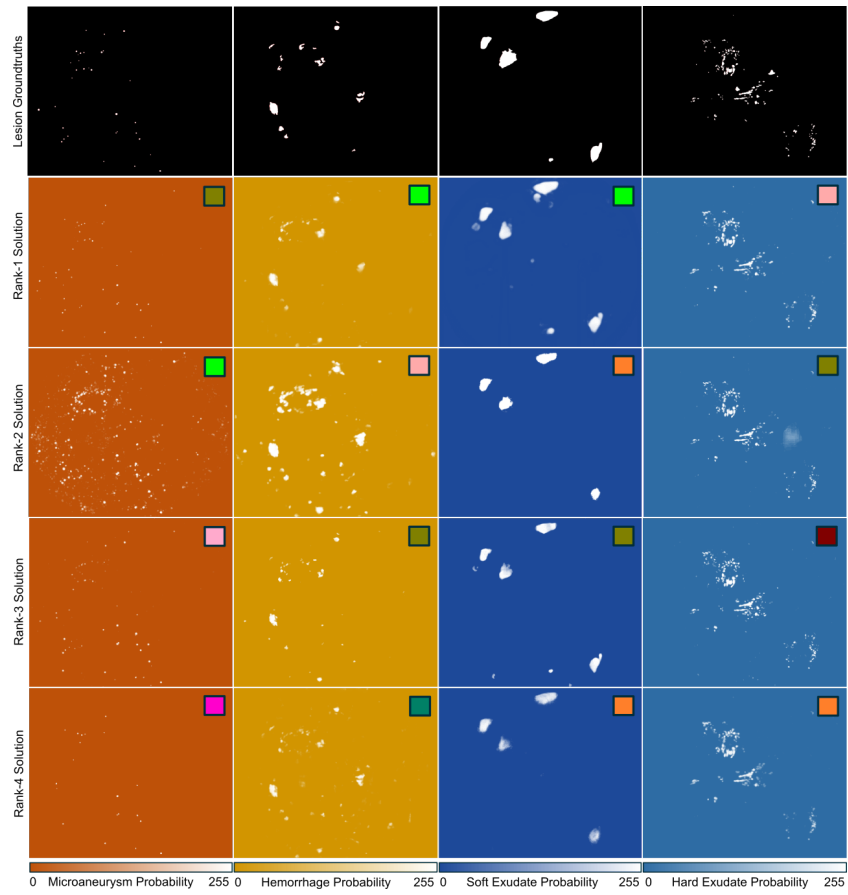
Lesion	Team Name	AUPR	Approach	Ensemble	Input Size (Pixels)	External Dataset
Microaneurysms	 iFLYTEK	0.5017	Cascaded CNN	✓	320 × 320	×
	 VRT	0.4951	U-Net	×	1280 × 1280	×
	 PATech	0.4740	DenseNet+U-Net	✓	256 × 256	×
	 SDNU	0.4111	Mask R-CNN	×	3584 × 2380	×
Hemorrhages	 VRT	0.6804	U-Net	×	640 × 640	×
	 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

1054 Table 8 summarizes the individual performance (Off-site evaluation) of each solu-
1055 tion listed in order of their final placement for each subtask. It also contains the various
1056 approaches followed and external dataset (if any) used for training the models. The
1057 higher the rank for individual task, the more favorable the performance. The top-3 en-
1058 tries according to the individual lesion segmentation task are VRT, iFLYTEK-MIG and
1059 PATech. Some sample lesion segmentation results illustrated in Fig. 6 and their corre-
1060 sponding overall evaluation score from Table 8 give a better idea of how the evaluation
1061 scores correlate with the quality of the segmentation.

1062 Fig. 7 summarizes the performance of top-4 teams per lesion segmentation task.
1063 The different curves represent the performance of the participating methods for various
1064 lesions (MAs, HEs, SEs and EXs). Team VRT achieved highest AUPR score for HE
1065 and SE segmentation task. Whereas, team PATech and iFLYTEK-MIG obtained best
1066 score for EX and MA segmentation task respectively.



(a)



(b)

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.

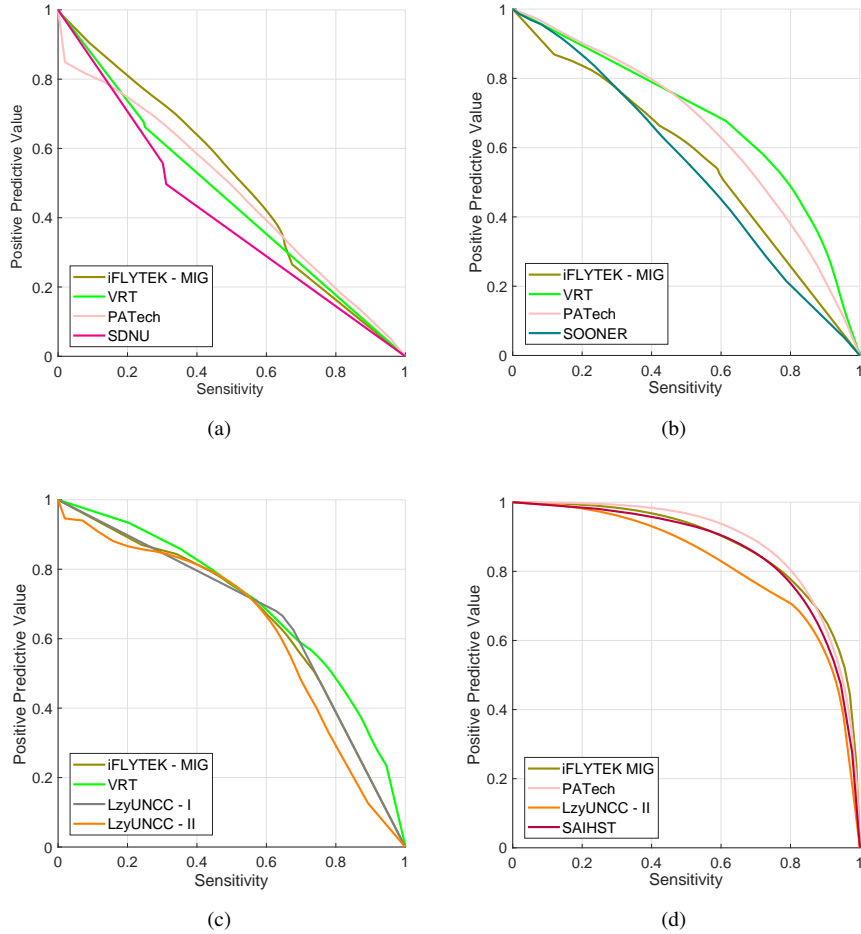


Fig. 7. The AUPR curves for the four top performing individual methods on the test dataset. These curves plot the sensitivity versus the positive predictive values for the different lesions, namely, (a) MAs, (b) HEs, (c) SEs, and (d) EXs

1067 *B. Sub-challenge – 2*

1068 This section presents the results achieved (On-site evaluation) by the participating
 1069 teams for the DR and DME grading task. It is important to note that this task was
 1070 evaluated for simultaneous grading of DR and DME using the validation algorithm
 1071 outlined in section 6.B on the test set (Set-B). This algorithm produced an average
 1072 grading accuracy of joint DR and DME on all images. Table 9 summarizes the result

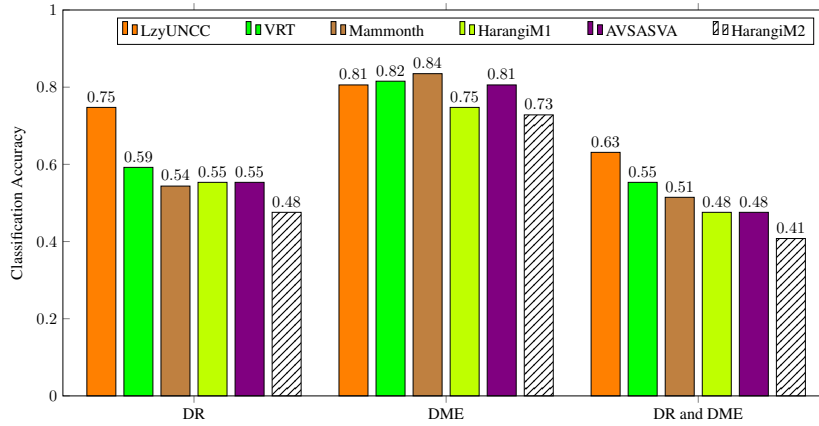


Fig. 8. Barplots showing separate and simultaneous classification accuracy of solutions developed by top - 6 teams for grading of DR and DME.

1073 of teams for on-site challenge along with the approach followed and the external dataset
 1074 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











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

1075 The top performing solution at the “on-site” challenge was proposed by team
 1076 LzyUNCC followed by team VRT and team Mammoth. Fig. 8 shows the average
 1077 accuracy of the competing solutions for individual as well as simultaneous for DR and
 1078 DME grading task. Teams are observed to perform poorly in the DR grading task that
 1079 reduced the overall accuracy for simultaneous grading of DR and DME. Major reason
 1080 seems to be the difficult test set, difficulty in accurately discriminating the DR severity
 1081 grades.

1082 *C. Sub-challenge – 3*

1083 This section presents the evaluation of “On-site” results for the participating teams
 1084 in the sub-challenge – 3, for all three subtasks. The results for subtasks of OD and
 1085 Fovea center localization were evaluated by euclidean distance, whereas for OD seg-
 1086 mentation results were evaluated and ranked using Jaccard similarity score as outlined
 1087 in section 6.C. Results from the on-site evaluations are reported in Table 10 and Table
 1088 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.

Localize	Team Name	ED	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	-

1089 The winning methods for the detection task were developed by team DeepDR and
 1090 team VRT, with DeepDR performing best in both OD and Fovea detection tasks. But
 1091 the winning entries for OD segmentation task were from teams ZJU-BII-SGEX, VRT
 1092 and IITKgpKLIV. Some sample OD segmentation results from these teams are illus-
 1093 trated in Fig. 9.

1094 Fig. 10 shows box-plots (McGill et al., 1978) illustrating the range of Euclidean
 1095 distances from the center of (a) optic disc and (b) fovea as well as (c) spread of Jaccard
 1096 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.

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	-

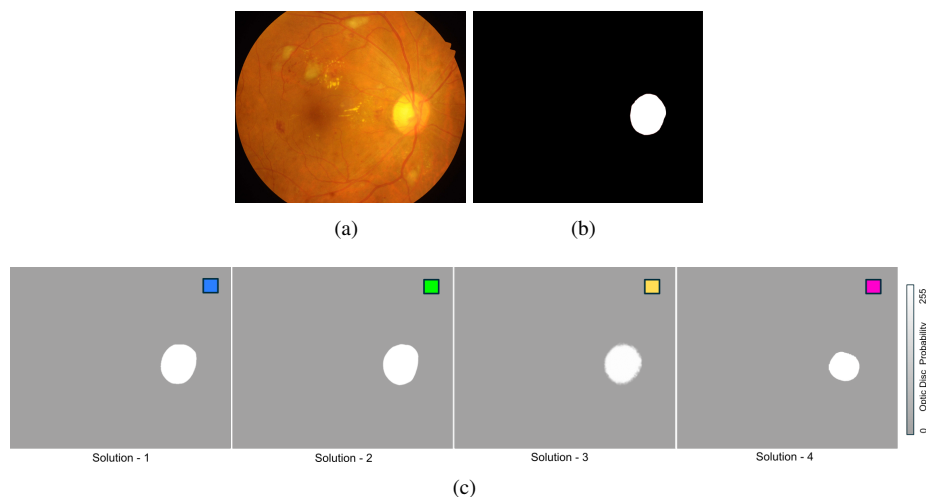


Fig. 9. Illustration of OD segmentation results: (a) sample image, (b) optic disc ground truth and (c) segmentation outcome of top-4 teams (from left to right)

1097 8. Discussion and Conclusion

1098 In this paper, we have presented the details of IDRiD challenge with detail infor-
 1099 mation about the data, evaluation metrics, an organization of the challenge, competing
 1100 solutions and final results for all sub-tasks, i.e., lesion segmentation, disease grading
 1101 and detection and segmentation of other normal retinal structures. Given the signifi-
 1102 cant number of participating teams (37) and results obtained, we believe this challenge
 1103 was a success. To the organizational end, efforts have been made in creating a rele-
 1104 vant, stimulating and fair competition, capable of advancing the collective knowledge

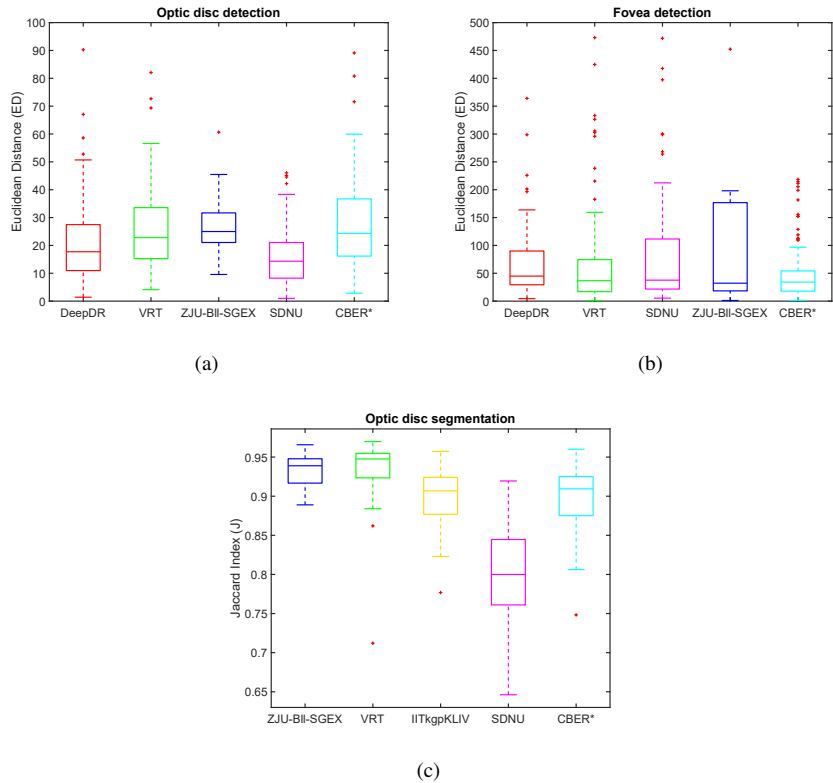


Fig. 10. Boxplots (a,b) showing dispersion of Euclidean distance for individual methods for OD and fovea and (c) showing the dispersion of Jaccard index for OD segmentation task. Boxplots show quartile ranges of the scores on the test dataset; plus sign indicate outliers (full range of data is not shown).

1105 in the research community. This section presents a discussion, limitations, and lessons
 1106 learned from this challenge.

1107 The first sub-challenge was conducted in an off-site mode in which 22 teams par-
 1108 ticipated with their lesion segmentation methods. The results of these methods on the
 1109 Set-B were evaluated by the organizers and amongst them, top-4 performing methods
 1110 per lesion segmentation task are included in this paper. The computed AUPR values
 1111 ranged between 0.4111 (for MAs) and 0.885 (for EXs). The best approach for lesion
 1112 segmentation used U-net, with data augmentation and the addition of dense block ex-
 1113 tract the features efficiently, boosting the results significantly. Fig. 11 highlights the
 1114 performance of top solution for EX that performs significantly well in presence of

normal retinal structures and different challenging circumstances. From the top per-

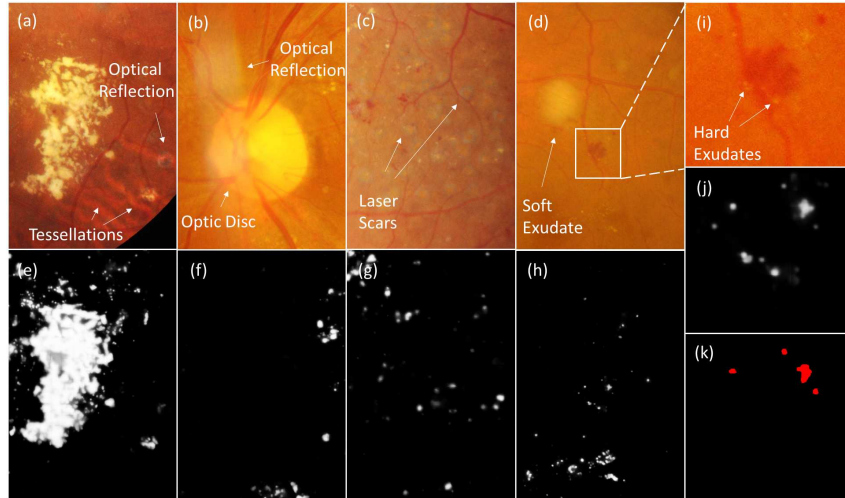


Fig. 11. Illustration of (a-d) different challenging circumstances for segmentation of EXs, (e-f) segmentation results (probability map) of top-performing team for EXs, (i) enlarged part of Fig. (d), and (j) depicts its performance to be better than (k) the human annotator (The annotator tool had limitation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

1115

1116 forming approaches, it is evident that solving the data imbalance problem improves
 1117 the model performance significantly. Since background overwhelms foreground, the
 1118 loss during training is more effectively back-propagated than that of foreground that
 1119 penalizes false negatives, boosting the sensitivity of lesion segmentation. Architectural
 1120 modifications to U-Net-based networks provided widely varying results for the differ-
 1121 ent types of lesion. For instance, the cascaded CNN approach yielded the best score
 1122 for MAs segmentation, as it add modules to reduce false positives. This approach dra-
 1123 matically impacts MA segmentation performance due the class imbalance of the task.
 1124 Further, Fig. 12 shows that some false positives detected by the participating solutions
 1125 are due to noise, predominantly for MA and HE. This indicates that there is still room
 1126 for improvement for lesion segmentation tasks with current fundus cameras.

1127

In the on-site disease-grading task six methods were compared and contrasted.
 1128 When assessed using the test data set hidden from the participants, the grading accu-
 1129 racy ranged between 0.4078 and 0.6311 as shown in Table 9. Notably, all teams
 1130 except AVASAVA used the external Kaggle DR dataset for pre-training their models.

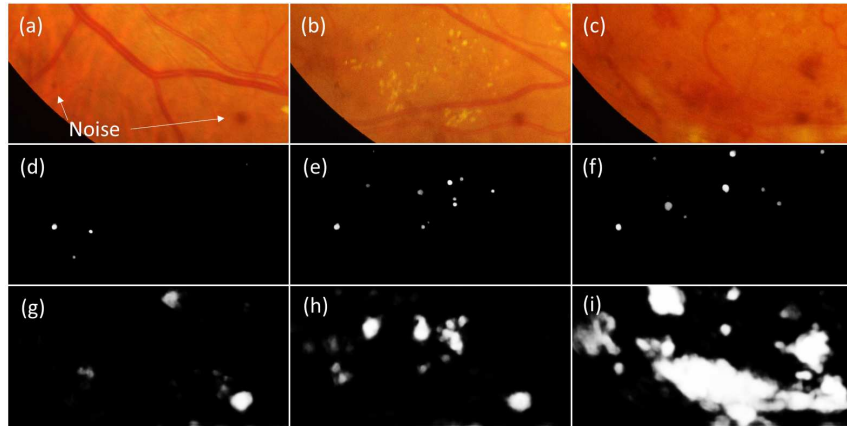


Fig. 12. Illustration of results by top performing solutions for (a-c) different images with noise causing most common false positives in the segmentation of (d-f) MAs, and (g-i) HEs respectively.

1131 This dataset contains a large amount of retina images annotated with the disease level,
 1132 in contrast, team AVASAVA pre-trained their model on ImageNet, a dataset containing
 1133 natural images and object annotations, effectively showing the network a much smaller
 1134 number of retina images at training stage, approximately 1% compared to the other
 1135 teams. This indicates that in the presence of a limited number of labeled data, transfer
 1136 learning approaches along with the good model pruning could yield comparable and
 1137 competitive results. However, while the models do determine the variability of the per-
 1138 formance, the number, type and quality of training data is a crucial factor for a fair
 1139 comparison of competing solutions. There is still work needed on simultaneous grad-
 1140 ing of DR and DME as the reported results do not yet reach the performance needed
 1141 for a clinically viable automatic screening. Considering the misclassified instances in
 1142 the confusion matrices in Table 12, along with the lesion information, it is essential
 1143 to give attention towards characterization of intra-retinal micro-vascular abnormalities
 1144 (IRMA's) and venous beading for improvement in the overall grading results.

1145 In the sub-challenge – 3, another on-site challenge, four teams were evaluated for
 1146 the task of OD/fovea localization and OD segmentation. For the task of OD localiza-
 1147 tion, the Euclidean Distance varied between 21.072 and 36.22 (lower values indicate
 1148 better performance). However, for Fovea localization task the same performance met-
 1149 ric ranged between 64.492 and 570.133. This massive variation is due to outliers, e.g.

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

		Predicted							Predicted		
		0	1	2	3	4			0	1	2
Actual	0	30	0	2	1	1	Actual	0	40	2	3
	1	3	1	1	0	0		1	5	2	3
	2	3	2	22	4	1		2	5	2	41
	3	2	0	1	13	3					
	4	1	0	1	0	11					

1150 team ZJU-BII-SGEX had 23 outliers whose Euclidean Distance exceeded 700. In the
 1151 OD segmentation task, the average Jaccard similarity index score amongst the partic-
 1152 ipants ranged between 0.7892 and 0.9338. The top-performing solutions developed
 1153 by DeepDR and VRT leveraged prior clinical knowledge, such as the number of land-
 1154 marks and their geometric relationship to detect another retinal landmark. It is also
 1155 observed that data augmentation and ensemble of models yield substantial improve-
 1156 ments in terms of accuracy. Considering the clinical significance of OD diameter while
 1157 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

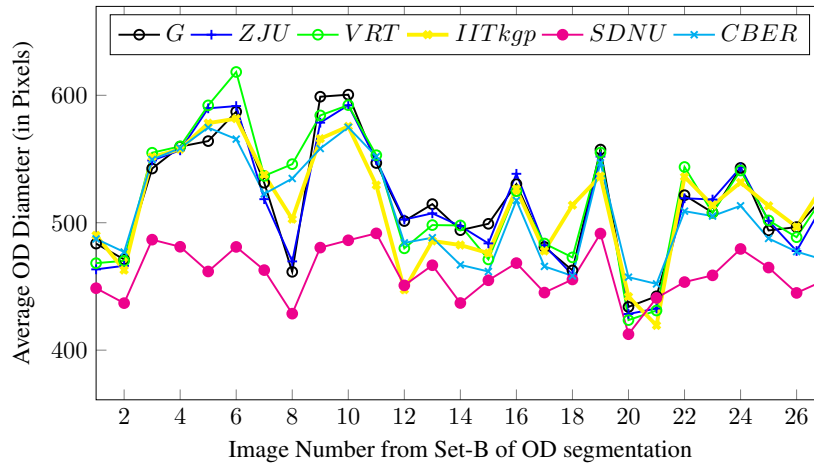


Fig. 13. Illustration of average OD diameter result of all 5 teams for each image of the testing dataset. [Here the legends G, ZJU and IITkgp represent Groundtruth, ZJU-BII-SGEX and IITkgpKLIV respectively (compressed to appear clearly in single column format, appears full in double column format.)]

1158

1159 with respect to the ground truth, most methods show a similar pattern. The average di-

1160 ameter of OD ground truth is 516.61 pixels whereas, this corresponding values for for
1161 the results of solutions developed by the teams ZJU-BII-SGEX, VRT, IITKgpKLIV,
1162 CBER and SDNU are 514.25, 519.21, 513.48, 508.04 and 460.19 pixels respectively.
1163 Team CBER submitted their after the competition and they were not included in the
1164 leaderboard.

1165 As expected, we found that image resolution is a vital factor for the model perfor-
1166 mance, especially for the task of segmentation of small objects such as MAs or EXs.
1167 In fact, the top performing approaches process the images patch-wise, which allow
1168 models to have a local high resolution image view or directly with the high resolu-
1169 tion image as a whole. This is essential as MAs or small EXs lesions span very few
1170 pixels in some cases, and reducing the original image size would prevent an accurate
1171 segmentation. Similarly, image resolution plays a very important role for the disease
1172 classification task (see Table 9), the most likely reason is that the presence of the dis-
1173 ease is determined by the presence of lesions in the image, including the small ones
1174 that might be invisible at low resolution. This is corroborated by the confusion matri-
1175 ces in Table 12 which show misclassified instances in DR (particularly, grade 1 and 2)
1176 as well as DME (5 images each belonging to grade 1 and 2 are predicted as grade 0).
1177 For the localization tasks, all participants were asked to identify retinal structures with
1178 coordinates at full image resolution. Most of them performed these tasks by scaling the
1179 image to smaller size and then converted their predictions in the original image space.
1180 The results indicate that the input image resolution has limited effect on the results of
1181 the localization problem. For instance, in case of OD localization, the top performing
1182 team utilized two image resolutions, one (224×224 pixels) for approximate location
1183 prediction and other (cropped ROIs 950×950 pixels) for refining that estimate. Sim-
1184 ilarly, teams CBER and VRT resized the image to 536×356 pixels and 640×640
1185 pixels respectively to get an approximate center location whereas, the team SDNU uti-
1186 lized the input size of 1984×1318 pixels. Considering the OD average diameter of
1187 approximately 516 pixels, the deflection of result for about 10 to 15 pixels by other
1188 approaches, utilizing different input resolutions, as compared to the top performing so-
1189 lution is very less. This is because the retinal structures to be identified, OD and fovea,
1190 are very unlikely to disappear due to a reduction of image resolution and they have

1191 clear geometrical constraints.

1192 This challenge provides data collected in the routine clinical practice and the ac-
1193 quisition protocol was consistent for all images. The data was acquired after pupil di-
1194 lation with the same camera at the same resolution, ensuring a consistent quality. This
1195 dataset did not include non-gradable images and images with substantial disagreement
1196 amongst the expert annotators. Even after these efforts to provide the best possible
1197 data, the annotation process is still inherently subjective and the annotator judgement is
1198 a limiting factor for the method performance which are mostly trained and evaluated in
1199 a supervised manner. While we believe that data challenges like ours foster “methodol-
1200 ogy diversity”, the majority of competing solutions used deep convolutional networks.
1201 These approaches are comparably easier to implement than approaches based on fea-
1202 ture engineering and do generalize well to multiple medical imaging domains dramati-
1203 cally reduces the need for specialized task knowledge. Notably, amongst the com-
1204 peting solutions in this challenge that utilized deep learning approach along with the
1205 task-relevant subject knowledge have demonstrated superior performance. However,
1206 it seems there might be some impact of challenge duration, apart from the number of
1207 submissions, on the quality of developed solutions. Considering the time span from
1208 data availability to deadline of results submission, about one and a half month, was
1209 considerably tight for managing all tasks at the same time. For the team VRT who
1210 had been working on analyzing fundus images for more than a year when participated
1211 in the competition that attempting all tasks were possible, still it was challenging for
1212 them to commit all the tasks. However, it would be highly challenging for a newcomer
1213 to succeed in multiple tasks. In that sense, the competition period was not sufficient
1214 for perfecting all tasks. However, it would be enough for a competent participant, e.g.
1215 new entrants in the field as team SAIHST, to finish one task if the participant can fo-
1216 cus on the competition completely. Also, in this challenge, the results were evaluated
1217 all at once after the result submission deadline. A continuous on-line assessment of
1218 participating solutions would have facilitated the submission procedure by providing
1219 real-time feedback to the team’s performance. This would have enabled a maximum
1220 number of submissions during the challenge period, probably boosting the final count
1221 of submissions. However, this would have introduced a risk of overfitting the test data

1222 by continuous submissions based on the system’s performance on the test set.

1223 This challenge led to the development of a variety of new robust solutions for le-
1224 sion segmentation, detection, and segmentation of retinal landmarks and disease sever-
1225 ity grading. Despite the complexity of the tasks, less than one-and-a-half month time
1226 for development, it received a very positive response, and the top performing solutions
1227 were able to achieve results close to the human annotators. Still, there is room for
1228 improvement, especially in the lesion segmentation and disease-grading tasks. Though
1229 the competition is now completed, the dataset has been made publicly available for re-
1230 search purposes to attract newcomers to the problem and to encourage the development
1231 of novel solutions to meet current and future clinical standards.

1232 **Acknowledgments**

1233 This work is sponsored by the Shri Guru Gobind Singhji Institute of Engineer-
1234 ing and Technology, Nanded (M.S.), INDIA. The authors would like to thank the fol-
1235 lowing people for their help in various aspects of organizing the ISBI-2018 Diabetic
1236 Retinopathy Segmentation and Grading Challenge: Prof. Emanuele Trucco (Univer-
1237 sity of Dundee, Scotland) and Tom MacGillivray (University of Edinburgh, Scotland),
1238 Ravi Kamble (SGGS Institute of Engineering and Technology, Nanded), Prof. Vivek
1239 Sahasrabudde (Government Medical College, Nanded) and Désiré Sidibé (Université
1240 de Bourgogne, France). We would also like to thank Prof. Jorge Cuadros, University
1241 of California, Berkeley (Organizer of Kaggle Diabetic Retinopathy challenge) for his
1242 kind permission for reporting the results of the models trained on their dataset.

1243 *VRT.* : This study was supported by the Research Grant for Intelligence Information
1244 Service Expansion Project, which is funded by National IT Industry Promotion Agency
1245 (NIPA-C0202-17-1045) in South Korea.

1246 *DeepDR.* : This work was supported in part by the National Natural Science Founda-
1247 tion of China under Grant Grant 61872241, Grant 61572316, in part by the National
1248 Key Research and Development Program of China under Grant 2016YFC1300302
1249 and Grant 2017YFE0104000, in part by the Science and Technology Commission of
1250 Shanghai Municipality under Grant 16DZ0501100 and Grant 17411952600.

1251 *HarangiM1-M2*. : Research was supported in part by the project EFOP-3.6.2-16-
1252 2017-00015 supported by the European Union and the State of Hungary, co-financed
1253 by the European Social Fund.

1254 *ZJU-BII-SGEX*. : This work is supported by Beijing Shangong Medical Technology
1255 Co., Ltd., which provided ocular healthcare solutions in China. Many thanks to the
1256 labeled images from Image Annotation Group of Beijing Shangong Medical Tech-
1257 nology.

1258 Team *CBER* (A.M. Mendonça, T. Melo, T. Araújo and A. Campilho) is financed by the
1259 ERDF – European Regional Development Fund through the Operational Programme
1260 for Competitiveness and Internationalisation - COMPETE 2020 Programme, and by
1261 National Funds through the FCT – Fundação para a Ciência e a Tecnologia (Portuguese
1262 Foundation for Science and Technology) within project CMUP-ERI/TIC/0028/2014.
1263 Teresa Araújo is funded by the FCT grant SFRH/BD/122365/2016.

1264 **Appendix A. Comparison of Publicly Available Retinal Image Databases**

1265 Table A.1 and Table A.2 provides the summary of technical specifications and avail-
1266 able ground truths in several existing datasets and the IDRiD dataset.

1267 **References**

1268 Abdulla, W., 2017. Mask r-cnn for object detection and instance segmentation on keras
1269 and tensorflow. https://github.com/matterport/Mask_RCNN.

1270 Abramoff, M. D., Garvin, M. K., Sonka, M., 2010. Retinal imaging and image analysis.
1271 IEEE reviews in biomedical engineering 3, 169–208.

1272 Abramoff, M. D., Lou, Y., Erginay, A., Clarida, W., Amelon, R., Folk, J. C., Niemei-
1273 jer, M., 2016. Improved automated detection of diabetic retinopathy on a publicly
1274 available dataset through integration of deep learning. Investigative ophthalmology
1275 & visual science 57 (13), 5200–5206.

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

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

EX - Hard Exudate, MA - Microaneurysms, H - Healthy, MY - Mydriatic, NMY - Non-Mydriatic, FOV - Field of View, LA - Lesion Annotation.

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

Name of Database	Normal Fundus Structures			Abnormalities				Multiple Experts		Disease Grading	Diabetic Macular Edema
	OD	VS	FA	MA	HE	EX	SE	Yes/No	#		
ARIA	✓	✓	✓	×	×	×	×	✓	2	×	×
DIARETDB1	×	×	×	✓	✓	✓	✓	✓	4	×	×
DRIVE	×	✓	×	×	×	×	×	✓	3	×	×
E-Ophtha	×	×	×	✓	×	✓	×	✓	2	×	×
HEIMED	×	×	×	×	×	✓	×	×	1	×	✓
Kaggle	×	×	×	×	×	×	×	✓	2	✓	×
MESSIDOR	×	×	×	×	×	×	×	×	1	✓	✓
ROC	×	×	×	✓	×	×	×	✓	4	×	×
STARE	✓	✓	×	×	×	×	×	✓	2	×	×
IDRiD	✓	×	✓	✓	✓	✓	✓	✓	2	✓	✓

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

1276 Acharya, R., Chua, C. K., Ng, E., Yu, W., Chee, C., 2008. Application of higher order
1277 spectra for the identification of diabetes retinopathy stages. Journal of Medical

1278 Systems 32 (6), 481–488.

1279 Acharya, U. R., Mookiah, M. R. K., Koh, J. E., Tan, J. H., Bhandary, S. V., Rao, A. K.,
1280 Hagiwara, Y., Chua, C. K., Laude, A., 2017. Automated diabetic macular edema
1281 (dme) grading system using dwt, dct features and maculopathy index. *Computers in*
1282 *biology and medicine* 84, 59–68.

1283 Acharya, U. R., Ng, E. Y.-K., Tan, J.-H., Sree, S. V., Ng, K.-H., 2012. An integrated
1284 index for the identification of diabetic retinopathy stages using texture parameters.
1285 *Journal of medical systems* 36 (3), 2011–2020.

1286 Adal, K. M., Sidibé, D., Ali, S., Chaum, E., Karnowski, T. P., Mériaudeau, F., 2014.
1287 Automated detection of microaneurysms using scale-adapted blob analysis and
1288 semi-supervised learning. *Computer methods and programs in biomedicine* 114 (1),
1289 1–10.

1290 Agurto, C., Murray, V., Barriga, E., Murillo, S., Pattichis, M., Davis, H., Russell, S.,
1291 Abràmoff, M., Soliz, P., 2010. Multiscale am-fm methods for diabetic retinopathy
1292 lesion detection. *IEEE transactions on medical imaging* 29 (2), 502–512.

1293 Almazroa, A., Alodhayb, S., Osman, E., Ramadan, E., Hummadi, M., Dlaim, M.,
1294 Alkatee, M., Raahemifar, K., Lakshminarayanan, V., 2018. Retinal fundus images
1295 for glaucoma analysis: the riga dataset. In: *Medical Imaging 2018: Imaging Inform-*
1296 *atics for Healthcare, Research, and Applications*. Vol. 10579. International Society
1297 for Optics and Photonics, p. 105790B.

1298 Antal, B., Hajdu, A., 2012. Improving microaneurysm detection using an optimally
1299 selected subset of candidate extractors and preprocessing methods. *Pattern Recogni-*
1300 *tion* 45 (1), 264–270.

1301 Antal, B., Hajdu, A., 2014. An ensemble-based system for automatic screening of
1302 diabetic retinopathy. *Knowledge-based systems* 60, 20–27.

1303 Atlas, I. D., 2017. Brussels, belgium: international diabetes federation. International
1304 Diabetes Federation (IDF).
1305 URL <http://diabetesatlas.org/resources/2017-atlas.html>

- 1306 Badrinarayanan, V., Kendall, A., Cipolla, R., 2015. Segnet: A deep convo-
1307 lutional encoder-decoder architecture for image segmentation. arXiv preprint
1308 arXiv:1511.00561.
- 1309 Bai, J., Miri, M. S., Liu, Y., Saha, P., Garvin, M., Wu, X., 2014. Graph-based optimal
1310 multi-surface segmentation with a star-shaped prior: Application to the segmenta-
1311 tion of the optic disc and cup. In: 2014 IEEE 11th International Symposium on
1312 Biomedical Imaging (ISBI). IEEE, pp. 525–528.
- 1313 Bandello, F., Parodi, M. B., Lanzetta, P., Loewenstein, A., Massin, P., Menchini, F.,
1314 Veritti, D., 2010. Diabetic macular edema. In: Macular Edema. Vol. 47. Karger Pub-
1315 lishers, pp. 73–110.
- 1316 Biyani, R., Patre, B., 2018. Algorithms for red lesion detection in diabetic retinopathy:
1317 A review. *Biomedicine & Pharmacotherapy* 107, 681–688.
- 1318 Bourne, R. R., Stevens, G. A., White, R. A., Smith, J. L., Flaxman, S. R., Price, H.,
1319 Jonas, J. B., Keeffe, J., Leasher, J., Naidoo, K., et al., 2013. Causes of vision loss
1320 worldwide, 1990–2010: a systematic analysis. *The lancet global health* 1 (6), e339–
1321 e349.
- 1322 Boyd, K., Eng, K. H., Page, C. D., 2013. Area under the precision-recall curve:
1323 Point estimates and confidence intervals. In: *Joint European Conference on Machine*
1324 *Learning and Knowledge Discovery in Databases*. Springer, pp. 451–466.
- 1325 Carin, L., Pencina, M. J., 2018. On deep learning for medical image analysis. *JAMA*
1326 320 (11), 1192–1193.
- 1327 Carmona, E. J., Rincón, M., García-Feijoó, J., Martínez-de-la Casa, J. M., 2008. Iden-
1328 tification of the optic nerve head with genetic algorithms. *Artificial Intelligence in*
1329 *Medicine* 43 (3), 243–259.
- 1330 Carson Lam, D. Y., Guo, M., Lindsey, T., 2018. Automated detection of diabetic
1331 retinopathy using deep learning. *AMIA Summits on Translational Science Proceed-*
1332 *ings* 2017, 147.

- 1333 Cheng, J., Yin, F., Wong, D. W. K., Tao, D., Liu, J., 2015. Sparse dissimilarity-
1334 constrained coding for glaucoma screening. *IEEE Transactions on Biomedical En-*
1335 *gineering* 62 (5), 1395–1403.
- 1336 Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way,
1337 G. P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M. M., et al., 2018. Opportu-
1338 nities and obstacles for deep learning in biology and medicine. *Journal of The Royal*
1339 *Society Interface* 15 (141), 20170387.
- 1340 Chudzik, P., Majumdar, S., Calivá, F., Al-Diri, B., Hunter, A., 2018. Microaneurysm
1341 detection using fully convolutional neural networks. *Computer methods and pro-*
1342 *grams in biomedicine* 158, 185–192.
- 1343 Ciulla, T. A., Amador, A. G., Zinman, B., 2003. Diabetic retinopathy and diabetic mac-
1344 ular edema: pathophysiology, screening, and novel therapies. *Diabetes care* 26 (9),
1345 2653–2664.
- 1346 Cuadros, J., Bresnick, G., 2009. Eyepacs: an adaptable telemedicine system for dia-
1347 betic retinopathy screening. *Journal of diabetes science and technology* 3 (3), 509–
1348 516.
- 1349 Dai, L., Fang, R., Li, H., Hou, X., Sheng, B., Wu, Q., Jia, W., 2018. Clinical re-
1350 port guided retinal microaneurysm detection with multi-sieving deep learning. *IEEE*
1351 *transactions on medical imaging* 37 (5), 1149–1161.
- 1352 Das, V., Puhan, N., Panda, R., 2015. Entropy thresholding based microaneurysm de-
1353 tection in fundus images. In: *2015 Fifth National Conference on Computer Vision,*
1354 *Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*. IEEE, pp. 1–4.
- 1355 Dashtbozorg, B., Mendonça, A. M., Campilho, A., 2015. Optic disc segmentation using
1356 the sliding band filter. *Computers in biology and medicine* 56, 1–12.
- 1357 Decencière, E., Cazuguel, G., Zhang, X., Thibault, G., Klein, J.-C., Meyer, F., Mar-
1358 cotegui, B., Quellec, G., Lamard, M., Danno, R., et al., 2013. Teleophtha: Machine
1359 learning and image processing methods for teleophthalmology. *IRBM* 34 (2), 196–
1360 203.

- 1361 Decencière, E., Zhang, X., Cazuguel, G., Laÿ, B., Cochener, B., Trone, C., Gain, P.,
1362 Ordóñez-Varela, J.-R., Massin, P., Erginay, A., et al., 2014. Feedback on a publicly
1363 distributed image database: the messidor database. *Image Analysis and Stereology*
1364 33 (3), 231–234.
- 1365 Deepak, K. S., Sivaswamy, J., 2012. Automatic assessment of macular edema from
1366 color retinal images. *IEEE Transactions on medical imaging* 31 (3), 766–776.
- 1367 Dhara, A. K., Mukhopadhyay, S., Bency, M. J., Rangayyan, R. M., Bansal, R., Gupta,
1368 A., 2015. Development of a screening tool for staging of diabetic retinopathy in
1369 fundus images. In: *Medical Imaging 2015: Computer-Aided Diagnosis*. Vol. 9414.
1370 International Society for Optics and Photonics, p. 94140H.
- 1371 Dobbin, K. K., Simon, R. M., 2011. Optimally splitting cases for training and testing
1372 high dimensional classifiers. *BMC medical genomics* 4 (1), 31.
- 1373 Esteves, T., Quelhas, P., Mendonça, A. M., Campilho, A., 2012. Gradient convergence
1374 filters and a phase congruency approach for in vivo cell nuclei detection. *Machine*
1375 *Vision and Applications* 23 (4), 623–638.
- 1376 Everingham, M., Van Gool, L., Williams, C. K., Winn, J., Zisserman, A., 2010. The
1377 pascal visual object classes (voc) challenge. *International journal of computer vision*
1378 88 (2), 303–338.
- 1379 Farnell, D. J., Hatfield, F., Knox, P., Reakes, M., Spencer, S., Parry, D., Harding, S.,
1380 2008. Enhancement of blood vessels in digital fundus photographs via the applica-
1381 tion of multiscale line operators. *Journal of the Franklin institute* 345 (7), 748–765.
- 1382 Ferris, F. L., 1993. How effective are treatments for diabetic retinopathy? *Jama*
1383 269 (10), 1290–1291.
- 1384 Figueiredo, I. N., Kumar, S., Oliveira, C. M., Ramos, J. D., Engquist, B., 2015. Auto-
1385 mated lesion detectors in retinal fundus images. *Computers in biology and medicine*
1386 66, 47–65.

- 1387 Fleming, A. D., Philip, S., Goatman, K. A., Olson, J. A., Sharp, P. F., 2006. Auto-
1388 mated microaneurysm detection using local contrast normalization and local vessel
1389 detection. *IEEE transactions on medical imaging* 25 (9), 1223–1232.
- 1390 Fraz, M., Badar, M., Malik, A., Barman, S., 2018. Computational methods for exudates
1391 detection and macular edema estimation in retinal images: a survey. *Archives of*
1392 *Computational Methods in Engineering*, 1–28.
- 1393 Fu, H., Cheng, J., Xu, Y., Wong, D. W. K., Liu, J., Cao, X., 2018. Joint optic disc
1394 and cup segmentation based on multi-label deep network and polar transformation.
1395 arXiv preprint arXiv:1801.00926.
- 1396 Fu, H., Xu, Y., Wong, D. W. K., Liu, J., 2016. Retinal vessel segmentation via deep
1397 learning network and fully-connected conditional random fields. In: *Biomedical*
1398 *Imaging (ISBI), 2016 IEEE 13th International Symposium on*. IEEE, pp. 698–701.
- 1399 García, G., Gallardo, J., Mauricio, A., López, J., Del Carpio, C., 2017. Detection of
1400 diabetic retinopathy based on a convolutional neural network using retinal fundus
1401 images. In: *International Conference on Artificial Neural Networks*. Springer, pp.
1402 635–642.
- 1403 Garcia, M., Sanchez, C. I., Poza, J., López, M. I., Hornero, R., 2009. Detection of
1404 hard exudates in retinal images using a radial basis function classifier. *Annals of*
1405 *biomedical engineering* 37 (7), 1448–1463.
- 1406 Gargeya, R., Leng, T., 2017. Automated identification of diabetic retinopathy using
1407 deep learning. *Ophthalmology* 124 (7), 962–969.
- 1408 Gegundez-Arias, M. E., Marin, D., Bravo, J. M., Suero, A., 2013. Locating the fovea
1409 center position in digital fundus images using thresholding and feature extraction
1410 techniques. *Computerized Medical Imaging and Graphics* 37 (5-6), 386–393.
- 1411 Giachetti, A., Ballerini, L., Trucco, E., 2014. Accurate and reliable segmentation of the
1412 optic disc in digital fundus images. *Journal of Medical Imaging* 1 (2), 024001.

- 1413 Giancardo, L., Meriaudeau, F., Karnowski, T. P., Li, Y., Garg, S., Tobin, K. W., Chaum,
1414 E., 2012. Exudate-based diabetic macular edema detection in fundus images using
1415 publicly available datasets. *Medical image analysis* 16 (1), 216–226.
- 1416 Giancardo, L., Meriaudeau, F., Karnowski, T. P., Li, Y., Tobin, K. W., Chaum, E.,
1417 2011. Microaneurysm detection with radon transform-based classification on retina
1418 images. In: 2011 Annual International Conference of the IEEE Engineering in
1419 Medicine and Biology Society. IEEE, pp. 5939–5942.
- 1420 Giancardo, L., Roberts, K., Zhao, Z., 2017. Representation learning for retinal vas-
1421 culature embeddings. In: *Fetal, Infant and Ophthalmic Medical Image Analysis*.
1422 Springer, pp. 243–250.
- 1423 Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedfor-
1424 ward neural networks. In: *Proceedings of the thirteenth international conference on*
1425 *artificial intelligence and statistics*. pp. 249–256.
- 1426 Greenspan, H., Van Ginneken, B., Summers, R. M., 2016. Guest editorial deep learning
1427 in medical imaging: Overview and future promise of an exciting new technique.
1428 *IEEE Transactions on Medical Imaging* 35 (5), 1153–1159.
- 1429 Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang,
1430 G., Cai, J., et al., 2018. Recent advances in convolutional neural networks. *Pattern*
1431 *Recognition* 77, 354–377.
- 1432 Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venu-
1433 gopalan, S., Widner, K., Madams, T., Cuadros, J., et al., 2016. Development and
1434 validation of a deep learning algorithm for detection of diabetic retinopathy in reti-
1435 nal fundus photographs. *Jama* 316 (22), 2402–2410.
- 1436 Harangi, B., Hajdu, A., 2014. Detection of exudates in fundus images using a marko-
1437 vian segmentation model. In: 2014 36th Annual International Conference of the
1438 IEEE Engineering in Medicine and Biology Society. IEEE, pp. 130–133.
- 1439 Hatanaka, Y., Nakagawa, T., Hayashi, Y., Kakogawa, M., Sawada, A., Kawase, K.,
1440 Hara, T., Fujita, H., 2008. Improvement of automatic hemorrhage detection meth-

1441 ods using brightness correction on fundus images. In: *Medical Imaging 2008:*
1442 *Computer-Aided Diagnosis*. Vol. 6915. International Society for Optics and Pho-
1443 tonics, p. 69153E.

1444 Havlicek, J. P., 1996. *Am-fm image models*. University of Texas at Austin.

1445 He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: *Computer Vision*
1446 *(ICCV), 2017 IEEE International Conference on*. IEEE, pp. 2980–2988.

1447 He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition.
1448 In: *Proceedings of the IEEE conference on computer vision and pattern recognition*.
1449 pp. 770–778.

1450 Hinton, G., 2018. Deep learning—a technology with the potential to transform health
1451 care. *JAMA* 320 (11), 1101–1102.

1452 Hoo-Chang, S., Roth, H. R., Gao, M., Lu, L., Xu, Z., Noguees, I., Yao, J., Mollura, D.,
1453 Summers, R. M., 2016. Deep convolutional neural networks for computer-aided de-
1454 tection: Cnn architectures, dataset characteristics and transfer learning. *IEEE trans-*
1455 *actions on medical imaging* 35 (5), 1285.

1456 Hoover, A., 1975. Stare database. Available: Available: <http://www.ces.clemson.edu/~ahoover/stare>.

1458 Howard, A. G., 2013. Some improvements on deep convolutional neural network based
1459 image classification. *arXiv preprint arXiv:1312.5402*.

1460 Iandola, F., Moskewicz, M., Karayev, S., Girshick, R., Darrell, T., Keutzer, K.,
1461 2014. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint*
1462 *arXiv:1404.1869*.

1463 ICO, 2017. *Guidelines for diabetic eye care, 2nd edn*. International Council of Oph-
1464 thalmology (ICO).

1465 Jaccard, P., 1908. Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci.*
1466 *Nat.* 44, 223–270.

- 1467 Javidi, M., Pourreza, H.-R., Harati, A., 2017. Vessel segmentation and microaneurysm
1468 detection using discriminative dictionary learning and sparse representation. *Com-*
1469 *puter methods and programs in biomedicine* 139, 93–108.
- 1470 Jelinek, H., Cree, M. J., 2009. Automated image detection of retinal pathology. *Crc*
1471 *Press*.
- 1472 Jonas, R. A., Wang, Y. X., Yang, H., Li, J. J., Xu, L., Panda-Jonas, S., Jonas, J. B.,
1473 2015. Optic disc-fovea distance, axial length and parapapillary zones. *the beijing*
1474 *eye study 2011. PloS one* 10 (9), e0138701.
- 1475 Jones, S., Edwards, R., 2010. Diabetic retinopathy screening: a systematic review of
1476 the economic evidence. *Diabetic medicine* 27 (3), 249–256.
- 1477 Jordan, K. C., Menolotto, M., Bolster, N. M., Livingstone, I. A., Giardini, M. E., 2017.
1478 A review of feature-based retinal image analysis. *Expert Review of Ophthalmology*
1479 12 (3), 207–220.
- 1480 Joshi, S., Karule, P., 2019. Mathematical morphology for microaneurysm detection in
1481 fundus images. *European Journal of Ophthalmology*, 1120672119843021.
- 1482 Kamble, R., Kokare, M., Deshmukh, G., Hussin, F. A., Mériaudeau, F., 2017. Local-
1483 ization of optic disc and fovea in retinal images using intensity based line scanning
1484 analysis. *Computers in biology and medicine* 87, 382–396.
- 1485 Kao, E.-F., Lin, P.-C., Chou, M.-C., Jaw, T.-S., Liu, G.-C., 2014. Automated detection
1486 of fovea in fundus images based on vessel-free zone and adaptive gaussian template.
1487 *Computer methods and programs in biomedicine* 117 (2), 92–103.
- 1488 Kauppi, T., Kamarainen, J.-K., Lensu, L., Kalesnykiene, V., Sorri, I., Uusitalo, H.,
1489 Kälviäinen, H., 2012. A framework for constructing benchmark databases and pro-
1490 tocols for retinopathy in medical image analysis. In: *International Conference on*
1491 *Intelligent Science and Intelligent Data Engineering*. Springer, pp. 832–843.
- 1492 Ker, J., Wang, L., Rao, J., Lim, T., 2018. Deep learning applications in medical image
1493 analysis. *IEEE Access* 6, 9375–9389.

- 1494 Khojasteh, P., Júnior, L. A. P., Carvalho, T., Rezende, E., Aliahmad, B., Papa, J. P.,
1495 Kumar, D. K., 2018. Exudate detection in fundus images using deeply-learnable
1496 features. *Computers in biology and medicine*.
- 1497 Kim, J., Hong, J., Park, H., Kim, J., Hong, J., Park, H., 2018. Prospects of deep learning
1498 for medical imaging. *Precision and Future Medicine* 2 (2), 37–52.
- 1499 Kingma, D. P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv
1500 preprint arXiv:1412.6980.
- 1501 Kollias, A. N., Ulbig, M. W., 2010. Diabetic retinopathy: early diagnosis and effective
1502 treatment. *Deutsches Arzteblatt International* 107 (5), 75.
- 1503 Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. Imagenet classification with deep
1504 convolutional neural networks. In: *Advances in neural information processing sys-*
1505 *tems*. pp. 1097–1105.
- 1506 Lam, C., Yu, C., Huang, L., Rubin, D., 2018. Retinal lesion detection with deep learn-
1507 ing using image patches. *Investigative ophthalmology & visual science* 59 (1), 590–
1508 596.
- 1509 Li, H., Chutatape, O., 2004. Automated feature extraction in color retinal images by a
1510 model based approach. *IEEE Transactions on biomedical engineering* 51 (2), 246–
1511 254.
- 1512 Lin, S., Ramulu, P., Lamoureux, E. L., Sabanayagam, C., 2016. Addressing risk fac-
1513 tors, screening, and preventative treatment for diabetic retinopathy in developing
1514 countries: a review. *Clinical & experimental ophthalmology* 44 (4), 300–320.
- 1515 Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van
1516 Der Laak, J. A., Van Ginneken, B., Sánchez, C. I., 2017. A survey on deep learning
1517 in medical image analysis. *Medical image analysis* 42, 60–88.
- 1518 Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for seman-
1519 tic segmentation. In: *Proceedings of the IEEE conference on computer vision and*
1520 *pattern recognition*. pp. 3431–3440.

- 1521 Lynch, S. K., Shah, A., Folk, J. C., Wu, X., Abramoff, M. D., 2017. Catastrophic fail-
1522 ure in image-based convolutional neural network algorithms for detecting diabetic
1523 retinopathy. *Investigative Ophthalmology & Visual Science* 58 (8), 3776–3776.
- 1524 Marin, D., Gegundez-Arias, M., Ponte, B., Alvarez, F., Garrido, J., Ortega, C., Vasallo,
1525 M., Bravo, J., 2018. An exudate detection method for diagnosis risk of diabetic
1526 macular edema in retinal images using feature-based and supervised classification.
1527 *Medical & biological engineering & computing* 56 (8), 1379–1390.
- 1528 Marin, D., Gegundez-Arias, M. E., Suero, A., Bravo, J. M., 2015. Obtaining optic
1529 disc center and pixel region by automatic thresholding methods on morphologically
1530 processed fundus images. *Computer methods and programs in biomedicine* 118 (2),
1531 173–185.
- 1532 Mary, M. C. V. S., Rajsingh, E. B., Jacob, J. K. K., Anandhi, D., Amato, U., Selvan,
1533 S. E., 2015. An empirical study on optic disc segmentation using an active contour
1534 model. *Biomedical Signal Processing and Control* 18, 19–29.
- 1535 McGill, R., Tukey, J. W., Larsen, W. A., 1978. Variations of box plots. *The American*
1536 *Statistician* 32 (1), 12–16.
- 1537 Medhi, J. P., Dandapat, S., 2014. Analysis of maculopathy in color fundus images. In:
1538 2014 Annual IEEE India Conference (INDICON). IEEE, pp. 1–4.
- 1539 Mendonca, A. M., Sousa, A., Mendonça, L., Campilho, A., 2013. Automatic localiza-
1540 tion of the optic disc by combining vascular and intensity information. *Computerized*
1541 *medical imaging and graphics* 37 (5-6), 409–417.
- 1542 Mookiah, M. R. K., Acharya, U. R., Chandran, V., Martis, R. J., Tan, J. H., Koh,
1543 J. E., Chua, C. K., Tong, L., Laude, A., 2015. Application of higher-order spectra
1544 for automated grading of diabetic maculopathy. *Medical & biological engineering &*
1545 *computing* 53 (12), 1319–1331.
- 1546 Mookiah, M. R. K., Acharya, U. R., Chua, C. K., Lim, C. M., Ng, E., Laude, A., 2013a.
1547 Computer-aided diagnosis of diabetic retinopathy: A review. *Computers in biology*
1548 *and medicine* 43 (12), 2136–2155.

- 1549 Mookiah, M. R. K., Acharya, U. R., Martis, R. J., Chua, C. K., Lim, C. M., Ng,
1550 E., Laude, A., 2013b. Evolutionary algorithm based classifier parameter tuning
1551 for automatic diabetic retinopathy grading: A hybrid feature extraction approach.
1552 Knowledge-based systems 39, 9–22.
- 1553 Morales, S., Engan, K., Naranjo, V., Colomer, A., 2017. Retinal disease screening
1554 through local binary patterns. IEEE journal of biomedical and health informatics
1555 21 (1), 184–192.
- 1556 Morales, S., Naranjo, V., Angulo, J., Alcañiz, M., 2013. Automatic detection of op-
1557 tic disc based on pca and mathematical morphology. IEEE transactions on medical
1558 imaging 32 (4), 786–796.
- 1559 Murphy, K. P., 2012. Machine learning: a probabilistic perspective. MIT press.
- 1560 Nagy, B., Harangi, B., Antal, B., Hajdu, A., 2011. Ensemble-based exudate detection
1561 in color fundus images. In: Image and Signal Processing and Analysis (ISPA), 2011
1562 7th International Symposium on. IEEE, pp. 700–703.
- 1563 Naqvi, S. A., Zafar, H. M., et al., 2018. Automated system for referral of cotton-wool
1564 spots. Current diabetes reviews 14 (2), 168–174.
- 1565 Niemeijer, M., Abramoff, M. D., Van Ginneken, B., 2009. Information fusion for di-
1566 abetic retinopathy cad in digital color fundus photographs. IEEE transactions on
1567 medical imaging 28 (5), 775–785.
- 1568 Niemeijer, M., Van Ginneken, B., Cree, M. J., Mizutani, A., Quellec, G., Sánchez,
1569 C. I., Zhang, B., Hornero, R., Lamard, M., Muramatsu, C., et al., 2010. Retinopathy
1570 online challenge: automatic detection of microaneurysms in digital color fundus
1571 photographs. IEEE transactions on medical imaging 29 (1), 185–195.
- 1572 Niemeijer, M., Van Ginneken, B., Staal, J., Suttorp-Schulten, M. S., Abramoff, M. D.,
1573 2005. Automatic detection of red lesions in digital color fundus photographs. IEEE
1574 Transactions on medical imaging 24 (5), 584–592.

- 1575 Nørgaard, M. F., Grauslund, J., 2018. Automated screening for diabetic retinopathy—a
1576 systematic review. *Ophthalmic research*.
- 1577 Orlando, J. I., Prokofyeva, E., del Fresno, M., Blaschko, M. B., 2018. An ensemble
1578 deep learning based approach for red lesion detection in fundus images. *Computer*
1579 *methods and programs in biomedicine* 153, 115–127.
- 1580 Osareh, A., Shadgar, B., Markham, R., 2009. A computational-intelligence-based ap-
1581 proach for detection of exudates in diabetic retinopathy images. *IEEE Transactions*
1582 *on Information Technology in Biomedicine* 13 (4), 535–545.
- 1583 Patton, N., Aslam, T. M., MacGillivray, T., Deary, I. J., Dhillon, B., Eikelboom, R. H.,
1584 Yogesana, K., Constable, I. J., 2006. Retinal image analysis: concepts, applications
1585 and potential. *Progress in retinal and eye research* 25 (1), 99–127.
- 1586 Perdomo, O., Ojalora, S., Rodríguez, F., Arevalo, J., González, F. A., 2016. A novel
1587 machine learning model based on exudate localization to detect diabetic macular
1588 edema.
- 1589 Pereira, C., Gonçalves, L., Ferreira, M., 2015. Exudate segmentation in fundus images
1590 using an ant colony optimization approach. *Information Sciences* 296, 14–24.
- 1591 Pereira, C. S., Mendonça, A. M., Campilho, A., 2007. Evaluation of contrast enhance-
1592 ment filters for lung nodule detection. In: *International Conference Image Analysis*
1593 *and Recognition*. Springer, pp. 878–888.
- 1594 Pires, R., Avila, S., Jelinek, H. F., Wainer, J., Valle, E., Rocha, A., 2017. Beyond
1595 lesion-based diabetic retinopathy: a direct approach for referral. *IEEE journal of*
1596 *biomedical and health informatics* 21 (1), 193–200.
- 1597 Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabudhe, V.,
1598 Meriaudeau, F., 2018a. Indian diabetic retinopathy image dataset (idrid). *IEEE Dat-*
1599 *aport*.
1600 URL <http://dx.doi.org/10.21227/H25W98>

- 1601 Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabudde, V.,
1602 Meriaudeau, F., 2018b. Indian diabetic retinopathy image dataset (idrid): A database
1603 for diabetic retinopathy screening research. *Data* 3 (3/25).
1604 URL <http://www.mdpi.com/2306-5729/3/3/25>
- 1605 Porwal, P., Pachade, S., Kokare, M., Giancardo, L., Mériaudeau, F., 2018c. Retinal im-
1606 age analysis for disease screening through local tetra patterns. *Computers in biology*
1607 *and medicine* 102, 200 – 210.
- 1608 Quellec, G., Charrière, K., Boudi, Y., Cochener, B., Lamard, M., 2017. Deep image
1609 mining for diabetic retinopathy screening. *Medical image analysis* 39, 178–193.
- 1610 Quellec, G., Lamard, M., Erginay, A., Chabouis, A., Massin, P., Cochener, B.,
1611 Cazuguel, G., 2016. Automatic detection of referral patients due to retinal patholo-
1612 gies through data mining. *Medical image analysis* 29, 47–64.
- 1613 Quellec, G., Lamard, M., Josselin, P. M., Cazuguel, G., Cochener, B., Roux, C.,
1614 2008. Optimal wavelet transform for the detection of microaneurysms in retina pho-
1615 tographs. *IEEE transactions on medical imaging* 27 (9), 1230–1241.
- 1616 Raman, R., Gella, L., Srinivasan, S., Sharma, T., 2016. Diabetic retinopathy: An epi-
1617 demic at home and around the world. *Indian journal of Ophthalmology* 64 (1), 69.
- 1618 Rangrej, S. B., Sivaswamy, J., 2017. Assistive lesion-emphasis system: an assistive
1619 system for fundus image readers. *Journal of Medical Imaging* 4 (2), 024503.
- 1620 Ravi, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., Yang,
1621 G.-Z., 2017. Deep learning for health informatics. *IEEE journal of biomedical and*
1622 *health informatics* 21 (1), 4–21.
- 1623 Reichel, E., Salz, D., 2015. Diabetic retinopathy screening. In: *Managing Diabetic Eye*
1624 *Disease in Clinical Practice*. Springer, pp. 25–38.
- 1625 Rocha, A., Carvalho, T., Jelinek, H. F., Goldenstein, S., Wainer, J., 2012. Points of
1626 interest and visual dictionaries for automatic retinal lesion detection. *IEEE transac-*
1627 *tions on biomedical engineering* 59 (8), 2244–2253.

- 1628 Romero-Oraá, R., Jiménez-García, J., García, M., López-Gálvez, M. I., Oraá-Pérez,
1629 J., Hornero, R., 2019. Entropy rate superpixel classification for automatic red lesion
1630 detection in fundus images. *Entropy* 21 (4), 417.
- 1631 Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for
1632 biomedical image segmentation. In: *International Conference on Medical image
1633 computing and computer-assisted intervention*. Springer, pp. 234–241.
- 1634 Roychowdhury, S., Koozekanani, D. D., Parhi, K. K., 2014. Dream: diabetic retinopa-
1635 thy analysis using machine learning. *IEEE journal of biomedical and health infor-
1636 matics* 18 (5), 1717–1728.
- 1637 Saito, T., Rehmsmeier, M., 2015. The precision-recall plot is more informative than the
1638 roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one* 10 (3),
1639 e0118432.
- 1640 Sánchez, C. I., García, M., Mayo, A., López, M. I., Hornero, R., 2009. Retinal image
1641 analysis based on mixture models to detect hard exudates. *Medical Image Analysis*
1642 13 (4), 650–658.
- 1643 Sánchez, C. I., Niemeijer, M., Išgum, I., Dumitrescu, A., Suttorp-Schulten, M. S.,
1644 Abràmoff, M. D., van Ginneken, B., 2012. Contextual computer-aided detection:
1645 Improving bright lesion detection in retinal images and coronary calcification iden-
1646 tification in ct scans. *Medical image analysis* 16 (1), 50–62.
- 1647 Seoud, L., Hurtut, T., Chelbi, J., Cheriet, F., Langlois, J. P., 2016. Red lesion detection
1648 using dynamic shape features for diabetic retinopathy screening. *IEEE transactions
1649 on medical imaging* 35 (4), 1116–1126.
- 1650 Shah, M. P., Merchant, S., Awate, S. P., 2018. Abnormality detection using deep neu-
1651 ral networks with robust quasi-norm autoencoding and semi-supervised learning.
1652 In: *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on.
1653 IEEE*, pp. 568–572.
- 1654 Shen, D., Wu, G., Suk, H.-I., 2017. Deep learning in medical image analysis. *Annual
1655 review of biomedical engineering* 19, 221–248.

- 1656 Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D.,
1657 Wang, Z., 2016. Real-time single image and video super-resolution using an efficient
1658 sub-pixel convolutional neural network. In: Proceedings of the IEEE Conference on
1659 Computer Vision and Pattern Recognition. pp. 1874–1883.
- 1660 Shortliffe, E. H., Blois, M. S., 2006. The computer meets medicine and biology: emer-
1661 gence of a discipline. In: Biomedical Informatics. Springer, pp. 3–45.
- 1662 Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale
1663 image recognition. arXiv preprint arXiv:1409.1556.
- 1664 Sivaswamy, J., Krishnadas, S., Joshi, G. D., Jain, M., Tabish, A. U. S., 2014. Drishti-
1665 gs: Retinal image dataset for optic nerve head (onh) segmentation. In: 2014 IEEE
1666 11th International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 53–56.
- 1667 Son, J., Bae, W., Kim, S., Park, S. J., Jung, K.-H., 2018. Classification of findings with
1668 localized lesions in fundoscopic images using a regionally guided cnn. In: Compu-
1669 tational Pathology and Ophthalmic Medical Image Analysis. Springer, pp. 176–184.
- 1670 Son, J., Park, S. J., Jung, K.-H., 2017. Retinal vessel segmentation in fundoscopic
1671 images with generative adversarial networks. arXiv preprint arXiv:1706.09318.
- 1672 Sopharak, A., Uyyanonvara, B., Barman, S., Williamson, T. H., 2008. Automatic de-
1673 tection of diabetic retinopathy exudates from non-dilated retinal images using math-
1674 ematical morphology methods. Computerized medical imaging and graphics 32 (8),
1675 720–727.
- 1676 Sreng, S., Maneerat, N., Hamamoto, K., Panjaphongse, R., 2019. Cotton wool spots
1677 detection in diabetic retinopathy based on adaptive thresholding and ant colony op-
1678 timization coupling support vector machine. IEEJ Transactions on Electrical and
1679 Electronic Engineering.
- 1680 Srivastava, R., Duan, L., Wong, D. W., Liu, J., Wong, T. Y., 2017. Detecting retinal
1681 microaneurysms and hemorrhages with robustness to the presence of blood vessels.
1682 Computer methods and programs in biomedicine 138, 83–91.

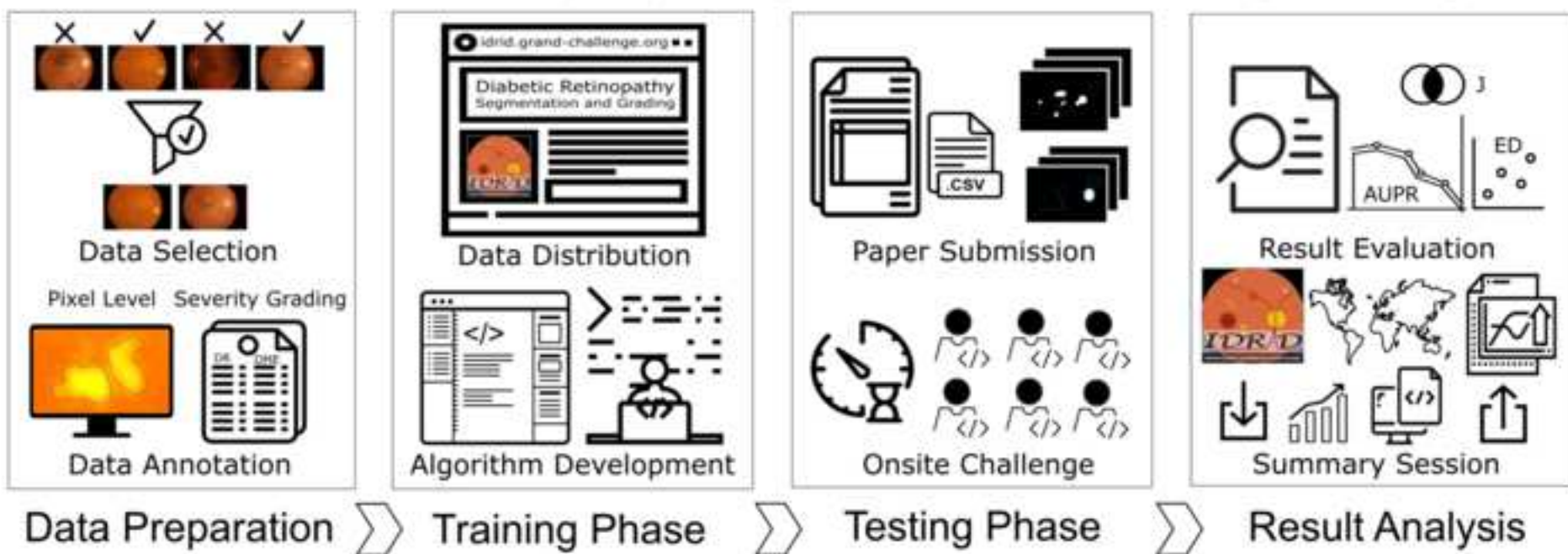
- 1683 Staal, J., Abramoff, M., Niemeijer, M., Viergever, M., van Ginneken, B., 2004. Ridge
1684 based vessel segmentation in color images of the retina. *IEEE Transactions on Med-*
1685 *ical Imaging* 23 (4), 501–509.
- 1686 Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M. J., 2017. Generalised
1687 dice overlap as a deep learning loss function for highly unbalanced segmentations.
1688 In: *Deep learning in medical image analysis and multimodal learning for clinical*
1689 *decision support*. Springer, pp. 240–248.
- 1690 Suzuki, K., 2017. Overview of deep learning in medical imaging. *Radiological physics*
1691 *and technology* 10 (3), 257–273.
- 1692 Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Van-
1693 houcke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: *Proceedings*
1694 *of the IEEE conference on computer vision and pattern recognition*. pp. 1–9.
- 1695 Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B.,
1696 Liang, J., 2016. Convolutional neural networks for medical image analysis: Full
1697 training or fine tuning? *IEEE transactions on medical imaging* 35 (5), 1299–1312.
- 1698 Tan, J. H., Acharya, U. R., Bhandary, S. V., Chua, K. C., Sivaprasad, S., 2017a. Seg-
1699 mentation of optic disc, fovea and retinal vasculature using a single convolutional
1700 neural network. *Journal of Computational Science* 20, 70–79.
- 1701 Tan, J. H., Fujita, H., Sivaprasad, S., Bhandary, S. V., Rao, A. K., Chua, K. C., Acharya,
1702 U. R., 2017b. Automated segmentation of exudates, haemorrhages, microaneurysms
1703 using single convolutional neural network. *Information sciences* 420, 66–76.
- 1704 Tang, L., Niemeijer, M., Reinhardt, J. M., Garvin, M. K., Abramoff, M. D., 2013.
1705 Splat feature classification with application to retinal hemorrhage detection in fundus
1706 images. *IEEE Transactions on Medical Imaging* 32 (2), 364–375.
- 1707 Thakur, N., Juneja, M., 2017. Clustering based approach for segmentation of optic cup
1708 and optic disc for detection of glaucoma. *Current Medical Imaging Reviews* 13 (1),
1709 99–105.

- 1710 Ting, D. S. W., Cheung, G. C. M., Wong, T. Y., 2016. Diabetic retinopathy: global
1711 prevalence, major risk factors, screening practices and public health challenges: a
1712 review. *Clinical & experimental ophthalmology* 44 (4), 260–277.
- 1713 Trucco, E., Ruggeri, A., Karnowski, T., Giancardo, L., Chaum, E., Hubschman, J. P.,
1714 Al-Diri, B., Cheung, C. Y., Wong, D., Abramoff, M., et al., 2013. Validating retinal
1715 fundus image analysis algorithms: issues and a proposal. *Investigative Ophthalmol-
1716 ogy & Visual Science* 54 (5), 3546–3559.
- 1717 Uribe-Valencia, L. J., Martínez-Carballido, J. F., 2019. Automated optic disc region
1718 location from fundus images: Using local multi-level thresholding, best channel
1719 selection, and an intensity profile model. *Biomedical Signal Processing and Control*
1720 51, 148–161.
- 1721 van Grinsven, M. J., van Ginneken, B., Hoyng, C. B., Theelen, T., Sánchez, C. I., 2016.
1722 Fast convolutional neural network training using selective data sampling: applica-
1723 tion to hemorrhage detection in color fundus images. *IEEE transactions on medical
1724 imaging* 35 (5), 1273–1284.
- 1725 Voulodimos, A., Doulamis, N., Bebis, G., Stathaki, T., 2018. Recent developments in
1726 deep learning for engineering applications. *Computational intelligence and neuro-
1727 science* 2018.
- 1728 Walter, T., Klein, J.-C., Massin, P., Erginay, A., 2002. A contribution of image process-
1729 ing to the diagnosis of diabetic retinopathy-detection of exudates in color fundus
1730 images of the human retina. *IEEE transactions on medical imaging* 21 (10), 1236–
1731 1243.
- 1732 Welfer, D., Scharcanski, J., Marinho, D. R., 2011. Fovea center detection based on the
1733 retina anatomy and mathematical morphology. *Computer methods and programs in
1734 biomedicine* 104 (3), 397–409.
- 1735 Winder, R. J., Morrow, P. J., McRitchie, I. N., Bailie, J., Hart, P. M., 2009. Algorithms
1736 for digital image processing in diabetic retinopathy. *Computerized medical imaging
1737 and graphics* 33 (8), 608–622.

- 1738 Wong, T. Y., Cheung, C. M. G., Larsen, M., Sharma, S., Simó, R., 2016. Diabetic
1739 retinopathy. *Nature Reviews Disease Primers*.
- 1740 Wu, B., Zhu, W., Shi, F., Zhu, S., Chen, X., 2017. Automatic detection of microa-
1741 neurysms in retinal fundus images. *Computerized Medical Imaging and Graphics*
1742 55, 106–112.
- 1743 Wu, L., Fernandez-Loaiza, P., Sauma, J., Hernandez-Bogantes, E., Masis, M., 2013.
1744 Classification of diabetic retinopathy and diabetic macular edema. *World journal of*
1745 *diabetes* 4 (6), 290.
- 1746 Wu, X., Dai, B., Bu, W., 2016. Optic disc localization using directional models. *IEEE*
1747 *Transactions on Image Processing* 25 (9), 4433–4442.
- 1748 Yu, F., Wang, D., Shelhamer, E., Darrell, T., 2017. Deep layer aggregation. *arXiv*
1749 *preprint arXiv:1707.06484*.
- 1750 Yu, H., Barriga, E. S., Agurto, C., Echegaray, S., Pattichis, M. S., Bauman, W., Soliz,
1751 P., 2012. Fast localization and segmentation of optic disk in retinal images using
1752 directional matched filtering and level sets. *IEEE Transactions on information tech-*
1753 *nology in biomedicine* 16 (4), 644–657.
- 1754 Yun, W. L., Acharya, U. R., Venkatesh, Y. V., Chee, C., Min, L. C., Ng, E. Y. K., 2008.
1755 Identification of different stages of diabetic retinopathy using retinal optical images.
1756 *Information sciences* 178 (1), 106–121.
- 1757 Zhang, B., Karray, F., Li, Q., Zhang, L., 2012. Sparse representation classifier for
1758 microaneurysm detection and retinal blood vessel extraction. *Information Sciences*
1759 200, 78–90.
- 1760 Zhang, X., Thibault, G., Decencière, E., Marcotegui, B., Laÿ, B., Danno, R., Cazuguel,
1761 G., Quellec, G., Lamard, M., Massin, P., et al., 2014. Exudate detection in color
1762 retinal images for mass screening of diabetic retinopathy. *Medical image analysis*
1763 18 (7), 1026–1043.

- 1764 Zhou, W., Wu, C., Yi, Y., Du, W., 2017. Automatic detection of exudates in digital
1765 color fundus images using superpixel multi-feature classification. *IEEE Access* 5,
1766 17077–17088.
- 1767 Zilly, J., Buhmann, J. M., Mahapatra, D., 2017. Glaucoma detection using entropy sam-
1768 pling and ensemble learning for automatic optic cup and disc segmentation. *Com-
1769 puterized Medical Imaging and Graphics* 55, 28–41.

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:

Reviewer #1:

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 arised 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; 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, 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

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

EX - Hard Exudate, MA - Microaneurysms, H - Healthy, MY - Mydriatic, NMY - Non-Mydriatic, FOV - Field of View, LA - Lesion Annotation.

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

Name of Database	Normal Fundus Structures			Abnormalities				Multiple Experts		Disease Grading	Diabetic Macular Edema
	OD	VS	FA	MA	HE	EX	SE	Yes/No	#		
ARIA	✓	✓	✓	×	×	×	×	✓	2	×	×
DIARETDB1	×	×	×	✓	✓	✓	✓	✓	4	×	×
DRIVE	×	✓	×	×	×	×	×	✓	3	×	×
E-Optha	×	×	×	✓	×	✓	×	✓	2	×	×
HEIMED	×	×	×	×	×	✓		×	1	×	✓
Kaggle	×	×	×	×	×	×	×	✓	2	✓	×
MESSIDOR	×	×	×	×	×	×	×	×	1	✓	✓
ROC	×	×	×	✓	×	×	×	✓	4	×	×
STARE	✓	✓	×	×	×	×	×	✓	2	×	×
IDRiD	✓	×	✓	✓	✓	✓	✓	✓	2	✓	✓

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

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-10 α 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 \times 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 ℓ^{th} layer, the feature value in the i^{th} feature map, $A_{x,y,i}^\ell$, is obtained as:

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

Where the parameters w_i^ℓ and b_i^ℓ are weight vector and bias term of the i^{th} filter of the 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 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 (Θ , 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 $\{(I^n, O^n); n \in [1, \dots, N]\}$ 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 ℓ layer network, instead of only ℓ , 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, \dots , 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.

Team Name	MA Score	RANK	HE Score	RANK	SE Score	RANK	EX Score	RANK
VRT	0.4951	2	0.6804	1	0.6995	1	0.7127	11
PATech	0.474	3	0.649	2	-	-	0.885	1
iFLYTEK-MIG	0.5017	1	0.5588	3	0.6588	3	0.8741	2
SOONER	0.4003	5	0.5395	4	0.5369	7	0.739	10
SAIHST	-	-	-	-	-	-	0.8582	3
lzyuncc_fusion	-	-	-	-	0.6259	4	0.8202	4
SDNU	0.4111	4	0.4572	7	0.5374	6	0.5018	17

➤ 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))

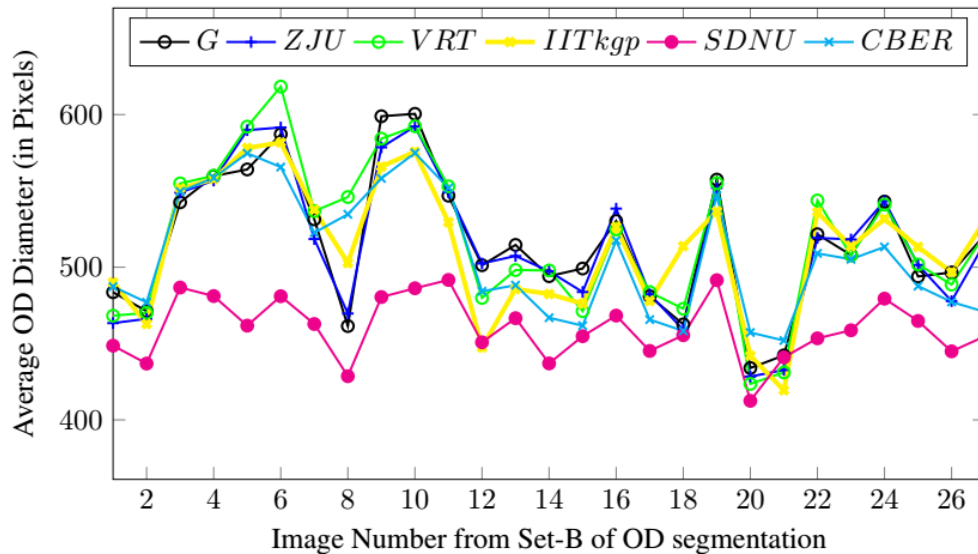


Fig. 13: Illustration of average OD diameter result of all 5 teams for each image of the testing dataset.[Here the legends G,ZJU and IITkqp represent Groundtruth, ZJU-BII-SGEX and IITKgpKLIV respectively (compressed to appear clearly in single column format, appears full in double column format.)]

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)

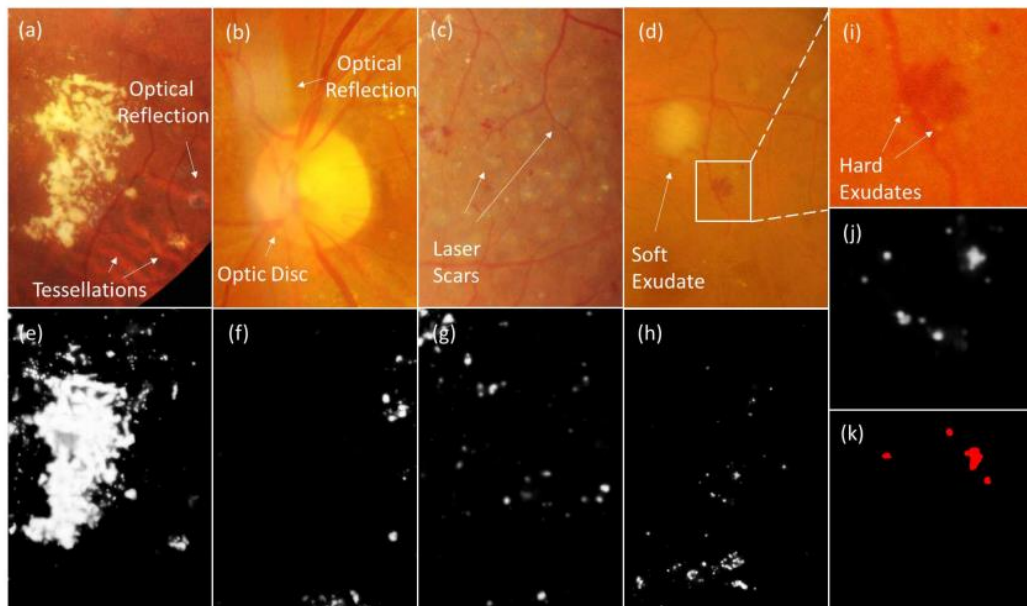


Fig. 11. Illustration of (a-d) different challenging circumstances for segmentation of EXs, (e-f) segmentation results (probability map) of top-performing team for EXs, (i) enlarged part of Fig. (d), and (j) depicts its performance to be better than (k) the human annotator (The annotator tool had limitation 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)

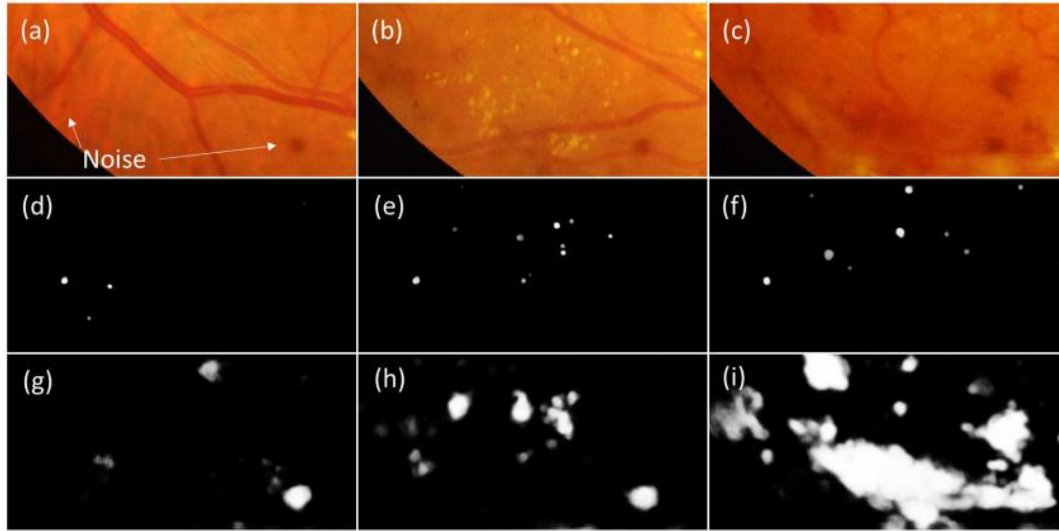


Fig. 12: Illustration of results by top performing solutions for (a-c) different images with noise causing most common false positives in the segmentation of (d-f) MAs, and (g-i) HEs respectively.

“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.” (line no. 1141-1144)

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

		Predicted				
		0	1	2	3	4
Actual	0	30	0	2	1	1
	1	3	1	1	0	0
	2	3	2	22	4	1
	3	2	0	1	13	3
	4	1	0	1	0	11

		Predicted		
		0	1	2
Actual	0	40	2	3
	1	5	2	3
	2	5	2	41

- 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.” (page no. 4 (line no. 25-32))

“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, Qin hao 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, Debdot 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)