Appendix B. Miscellaneous

1. Additional Results on External Datasets

We evaluate the different models on the test set for the Kaggle dataset and the Messidor-2 dataset. We show in table 1 the results of the five-class grading problem for DR on the external datasets. We see that the ensemble of three EfficientNet models is doing the best. Also, the best standalone model has the EfficientNet backbone. We show in tables 2 and 3 the results on the binary referability and vision-threatening problems. It is obvious that the ensemble of EfficientNet models is outperforming other models. It is to be noted that the models were trained only on the Kaggle training dataset. The results we obtained on the Messidor-2 dataset confirm the generalization of the trained models.

Based on the results shown in tables 1, 2 and 3, we decided to train only EfficientNet-based models for DR and DME existence classification on the DR10K dataset.

	Kaggle Test		Messidor-2	
	Accuracy	QWK	Accuracy	QWK
1. Inception-V3-small+lr 5e-5	86.01	82.38	80.22	86.73
2. Densenet-161-small+lr 1e-5	86.36	82.27	77.35	82.36
3. Eff-b5-small+lr 1e-4	86.59	83.18	80.1	87.94
4. Eff-b7-small+lr 1e-4	**87.06**	83.65	80.28	87.26
5. Eff-b7 Big +lr 1e-04	86.97	**83.89**	**82.86**	88.22
6. Mobilenet-V2-small+lr 1e-4	86.29	83.06	80.91	86.31
7. Resnet152-small+lr 1e-4	86.51	82.86	80.16	86.47
8. Vit-small+lr 1e-5	85.1	79.7	75.06	79.83
9. Convnext-Big+lr 1e-5	84.36	79.25	80.22	87.41
10. Swin-small+lr 5e-5	86.8	83.58	80.68	**88.56**
Ensemble of 1,2 and 5	87.33	84.05	82.51	88.62
Ensemble of 3,4 and 5	**87.69**	**84.8**	**83.6**	**90.06**

Table 1. Results for five-class DR grading problem on the Kaggle test set and the Messidor dataset.

	Kaggle Test Set for binary DR referability problem				
	Acc.	Sens.	Spec.	H-Mean	AUC
1. Inception-V3-small+lr 1e-4	93.6	78.75	97.05	86.94	95.51
2. Densenet-161-small+lr 1e-5	93.54	**83.03**	95.99	**89.04**	95.56
3. Eff-b5-small+lr 1e-4	93.75	81.32	96.63	88.32	96.16
4. Eff-b7-small+lr 5e-5	94.27	80.69	97.42	88.27	96.25
5. Eff-b7 Big +lr 1e-04	**94.34**	77.43	**98.27**	86.62	**96.29**
6. Mobilenet-V2-small+lr 1e-4	93.57	77.72	97.25	86.4	95.89
7. Resnet152-small+lr 1e-4	93.86	79.91	97.1	87.67	95.97
8. Vit-small+lr 1e-5	93.11	76.02	97.08	85.27	94.34
9. Convnext-Big+lr 1e-5	92.47	63.52	99.2	77.45	95.01
10. Swin-small+lr 5e-5	93.98	81.18	96.96	88.37	95.71
Ensemble of 1,2 and 5	94.37	80.64	97.56	88.3	96.32
Ensemble of 3,4 and 5	**94.59**	**81.08**	**97.73**	**88.63**	**96.55**
	Kaggle Test Set for binary DR VT problem				
1. Inception-V3-small+lr 5e-5	97.38	**68.97**	98.68	**81.19**	97.14
2. Densenet-161-small+lr 1e-5	97.53	59.33	99.28	74.27	97.14
3. Eff-b5-small+lr 1e-4	97.46	65.41	98.93	78.75	97.4
4. Eff-b7-small+lr 5e-5	97.33	65.62	98.78	78.86	**97.44**
5. Eff-b7 Big +lr 1e-04	97.59	61.64	99.23	76.04	97.43
6. Mobilenet-V2-small+lr 1e-4	**97.63**	61.22	99.3	75.74	97.36
7. Resnet152-small+lr 1e-4	97.37	61.84	98.99	76.13	97.17
8. Vit-small+lr 1e-5	97.49	57.65	99.31	72.95	96.88
9. Convnext-Big+lr 1e-5	97.29	44.65	**99.69**	61.68	96.74
10. Swin-small+lr 5e-5	**97.63**	60.38	99.34	75.11	97.18
Ensemble of 1,2 and 5	**97.72**	64.78	**99.22**	78.38	97.56
Ensemble of 3,4 and 5	97.62	**64.99**	99.11	**78.5**	**97.6**

Table 2. Results for binary DR problems on Kaggle dataset.

	Messidor-2 Test Set for binary DR referability problem				
	Acc.	Sens.	Spec.	H-Mean	AUC
1. Inception-V3-small+lr 1e-4	92.49	78.99	97.28	87.19	97.58
2. Densenet-161-small+lr 1e-5	89.16	**92.78**	87.88	90.26	97.2
3. Eff-b5-small+lr 1e-4	92.89	83.59	96.19	89.45	97.88
4. Eff-b7-small+lr 5e-5	92.66	89.06	93.94	91.43	97.7
5. Eff-b7 Big +lr 1e-04	**94.04**	90.37	95.34	**92.79**	**98.2**
6. Mobilenet-V2-small+lr 1e-4	92.32	87.31	94.09	90.57	97.8
7. Resnet152-small+lr 1e-4	91.11	87.96	92.23	90.05	97.9
8. Vit-small+lr 1e-5	92.32	82.93	95.65	88.84	95.9
9. Convnext-Big+lr 1e-5	93.58	79.43	**98.6**	87.98	98.06
10. Swin-small+lr 5e-5	92.55	**92.78**	92.46	92.62	97.62
Ensemble of 1,2 and 5	93.69	89.5	95.18	92.25	98.24
Ensemble of 3,4 and 5	**94.55**	**91.03**	**95.8**	**93.36**	**98.3**
	Messidor-2 Test Set for binary DR VT problem				
1. Inception-V3-small+lr 5e-5	97.48	67.27	99.51	80.28	98.94
2. Densenet-161-small+lr 1e-5	96.96	58.18	99.57	73.45	98.75
3. Eff-b5-small+lr 1e-4	97.31	63.64	99.57	77.65	98.9
4. Eff-b7-small+lr 5e-5	97.36	**90.91**	97.8	**94.23**	**99.16**
5. Eff-b7 Big +lr 1e-04	**97.88**	87.27	98.59	92.59	98.89
6. Mobilenet-V2-small+lr 1e-4	97.48	61.82	**99.88**	76.37	99.03
7. Resnet152-small+lr 1e-4	97.42	70	99.27	82.1	99.04
8. Vit-small+lr 1e-5	95.87	41.82	99.51	58.89	98.52
9. Convnext-Big+lr 1e-5	97.13	57.27	99.82	72.78	99
10. Swin-small+lr 5e-5	97.71	76.36	99.14	86.28	98.99
Ensemble of 1,2 and 5	97.71	73.64	**99.33**	84.57	99.13
Ensemble of 3,4 and 5	**98.22**	**87.27**	98.96	**92.75**	**99.2**

Table 3. Results for binary DR problems on Messidor dataset.

2. Comparison between the 3 used datasets

In table 4 we provided a comprehensive comparison between all the used datasets from different perspectives. This comparison highlights some strength points for the new DR10K dataset compared to the publicly available ones. DR10K is the only dataset that contains the papilla centered images in addition to the macula centered ones which enables us to augment the training using them enhancing the results as shown in the manuscript. While Kaggle dataset is only annotated for the DR 5 levels problem, Messidor-2 and DR10K is additionally annotated for the DME existence and image gradability problems. The advantage of DR10K is the presence of more than 1000 non-gradable images compared to only 4 images in messidor-2. Moreover, DR10K is the only dataset whose demographic data such as age, blood sugar level, diabetes type and duration available. The purpose of each dataset is relevant to its size so, Kaggle the largest one is used for training, DR10K which is mid-size is needed for finetuning as shown in the manuscript and Messidor-2 can only be used for testing to prove generalization due to its small size.

	Kaggle	**Messidor-2**	**DR10K**
Purpose	Training	Testing	Finetuning
National	America / India	France	Egypt
Cohort (Collection Source)	Community-based, Clinic-based	Clinic-based	Community-based, Population-based, Clinic-based
Camera	Variety of Types	Topcon	Optomed Aurora
Annotation	Not Mentioned	3 Graders, Discuss to solve conflicts	3 Graders , 1 Adjudicator solve conflicts
Annotated Problems	DR	DR, DME and Gradability	DR, DME and Gradability
Fundus Image Types	Macula Centered Only	Macula Centered Only	Both Macula and Papilla Centered
Demographic Data Availability	No	No	Yes
No. of Patients	N/A	874	11109
No. of Eyes	92363	1748	20387
No. of Images	92363	1748	40774

Table 4. Comprehensive comparison between Kaggle, Messidor-2 and DR10K datasets.

3. Datasets DR levels distribution

In figure 1 we show the distribution of our 3 used datasets w.r.t the DR levels. We can notice that the 3 distributions are not identical. The most important difference is that level 2 images are more than that of level 1 in both Kaggle and Messidor-2 while in DR10K the situation is reversed. This can highlight again the importance of collecting an Egyptian dataset to reflect the local distribution.

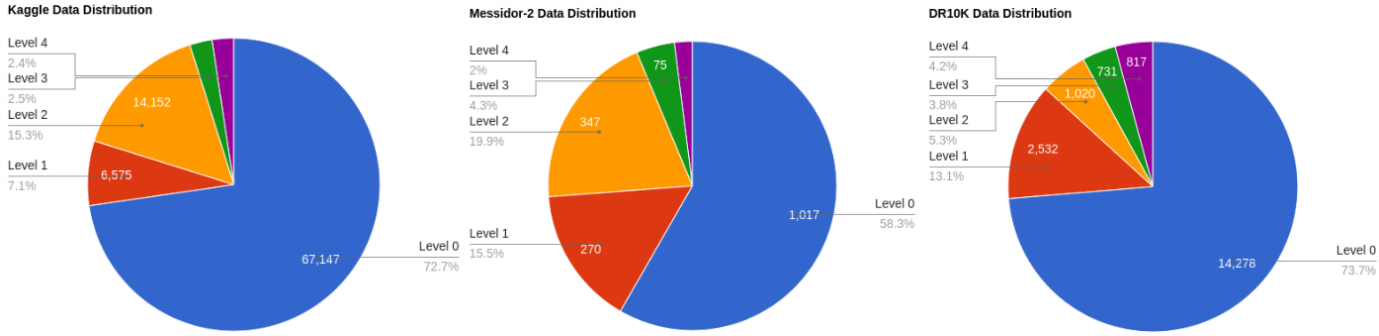

Figure 1. Per level Kaggle, Messidor-2 and DR10K datasets distribution