DR10K: Transfer Learning Using Weak Labels for Grading Diabetic Retinopathy on DR10K Dataset

Mohamed ElHabebe[1,2] Shereen ElKordi[2] Ahmed Gamal ElDin[2]

Noha Adly[1,2] Marwan Torki[1,2] Ahmed Elmassry[1] Islam SH Ahmed[1]

1.Alexandria University, Alexandria, Egypt
2.Applied Innovation Center, MCIT, Cairo, Egypt

Abstract

In this paper, we contrast the usage of two deep-learning approaches for the automatic grading of diabetic retinopathy (DR) and diabetic macular edema (DME) in retinal fundus photographs using a relatively small novel dataset. We developed a telemedicine system to collect and humanly grade 11,109 diabetic patients. The certified graders annotated the level of DR as well as the existence of a referable DME in the macula-centered fundus images only. We use EfficientNet to build an AI-based model for both problems. To examine the transfer learning validity, the model was trained on an external dataset (EyePacs) and then fine-tuned on the egyptian data for the DR and DME grading problems. Firstly, we use the macula-centered images only in fine-tuning. Secondly, we use optic-disc-centered images in addition to macula-centered images. We obtained the labels for the optic-disc-centered images directly from the corresponding macula-centered labels as weak labels. Then, both types of images are used in fine-tuning. We found an increase in the DR performance using the second approach in both accuracy and quadratic weighted kappa(QWK). Notably, QWK increased from 90.23% to 91.3% using additional weakly labeled optic-disc-centered fundus images.

1. Introduction

Diabetes is one of the most widely spread diseases around the world, especially in many developing countries. According to the international diabetes federation, approximately 463 million adults (20-79 years) are living with diabetes throughout the world [47].

Diabetic Retinopathy is one of the most dangerous complications of diabetes as it affects eyesight and may cause blindness [29, 49]. People who have had diabetes for many years (more than 15) are most probably suffering from diabetic retinopathy, but the early detection of this complication helps a lot in avoiding its development to severe levels [48] [2]. Diabetic Macular edema is another dangerous complication [6] which represents a sign of diabetic retinopathy evolution to the worst levels. Fundus images have proved high accuracy in allowing ophthalmologists to detect the presence of DR in addition to its severity level and the presence of DME [15, 54].

According to the proposed international clinical diabetic retinopathy disease severity scales (PIRC), fundus images are graded from 0 to 4 as follows:{ R0: no DR, R1: mild DR, R2: moderate DR, R3: severe DR and R4: proliferative DR}

We can also classify the DR cases into referable (RDR) and non-referable (NRDR) where referable cases are the cases that need to visit the doctor. These are the cases with classes R2, R3, and R4 on the PIRC scale. Moreover, the cases can be classified into vision-threatening and non-vision-threatening DR where the vision-threatening class includes DR levels R3 and R4. These cases may need urgent surgery to avoid eyesight loss [49]. Also, the detection of the presence of diabetic macular edema is an important sign that the patient needs to visit the doctor.

Deep learning-based models are widely used in the DR classification problem from fundus images [3, 28, 66]. They are mostly employed in screening programs to help in the early detection and regular follow-up of DR for diabetes patients. The screening program may be automated where the patient is directed to an ophthalmic clinic if classified as having referable DR by the model. It also may be semi-automated where the fundus image, classified by the model as referable, is manually reclassified by a human grader [59]. In both cases, if the patient is classified by

the model as having non-referable DR, he will redo the test after 6 months or a year. These programs save much time for the ophthalmologists which they were wasting on examining the non-referable DR cases that do not need medical intervention.

With millions of our population suffering from diabetes, it is important to develop a DR screening program. We can use publicly available DR classification datasets such as EyePacs and APTOS datasets which were released in 2015 and 2019 respectively and available on the Kaggle website to train our classification models. Unfortunately, the classification accuracy degrades noticeably when a model is used to classify fundus images from different ethnic groups than that of the training data. This may happen due to various factors such as the difference in the data distribution which will be shown as a crucial factor but not the only one in our results. The difference in the used fundus camera is also a factor in addition to biological differences as mentioned in [49]. To our knowledge, there is no available large DR classification dataset collected from Egypt. So, we collected a new dataset which is graded according to the DR PIRC scale and also for the presence or absence of DME.

This paper studied several dimensions to reach a complete pipeline for training the deep learning models needed in a DR screening program. The prevalence of DR in patients with diabetes in Egypt has been studied in [14]. Our contributions can be summarized as follows:

- Collect the DR10K dataset, grade it for DR and DME classification, and analyze it.
- Fine-tune best models trained using publicly available DR datasets on the DR10K dataset to classify DR and DME, achieving an accuracy of 89.51% in the 5-class DR classification.
- Using weakly labeled optic-disc centered images to augment the data of macula centered ones, achieving a better performance of 89.86% in the 5-class accuracy.
- Apply regression score thresholding to improve the classification results on the three binary problems: DME, DR referability, and DR vision threatening. Regression score thresholding shows great improvement in binary problems with respect to sensitivity.

2. Related Work

Convolutional Neural Networks (CNN) [30] are considered a breakthrough in image processing. It has the ability to learn different features starting from low level and going to mid-level and high-level complex features. Hence, it was indulged in many applications such as in image classification [32], age and gender estimation [33], image recognition [41], image segmentation [20, 34, 65], object detection [11] and video segmentation [37]. Many attempts have been done to interpret the learned features in these different problems [13, 51]

Semi-supervised learning or weakly supervised learning [69] is widely used to solve the problem of the lack of large annotated datasets for various machine learning problems including computer vision [16, 38, 42, 68]. Yalniz et al. [67] proposed a pipeline based on a teacher-student paradigm to allow the use of a billion of unlabelled images to enhance the performance of different CNN architectures achieving 81.2% top-1 accuracy on Imagenet using vanilla ResNet-50. Weak supervision helps especially in solving some highlighted challenges in medical field problems such as the high cost and effort of data annotation and the class imbalance [21, 23, 45].

Deep learning [31] has shown much potential in solving several problems in the field of computer vision, especially in the medical field. There have been attempts to detect many types of cancer such as breast cancer [39], skin cancer [7], and brain tumor classification [10]. One of these applications is eye disease classification such as glaucoma detection [64]. Diabetic retinopathy is one of those critical diseases. Many attempts to solve the diabetic retinopathy classification problem have been made. [50, 61, 62] used Inception-v3 [57] architecture to build classifiers for diabetic retinopathy and macular edema. Wang et al. [63] proposed two convolutional neural network structures and used regression activation maps to localize features that are discriminative and that contribute to the classification result. More recent works [25, 55] employed vision transformers in DR classification.

Classifiers were also built to detect diabetic macular edema in [8]. They used an ensemble of AlexNet [27], VGGNet [53], and GoogleNet [56]. In [61] they used an Inception-v3 [57] model to predict DME and predict a subretinal and an intraretinal fluid presence in a multi-task manner. They trained the model on fundus images while the labels were extracted from their OCT counterpart. The model showed better results than manual grading and they proved the model extracted DME-related features mainly from the optic disk when training using crops of the image. Also, [24] used vision transformer-based models to effectively classify DME from OCT images.

Due to diabetic retinopathy's dangerous consequences, many countries have launched screening programs to aid in the early detection and tracking of patient's history and treatment [43]. In Portugal, they achieved a sensitivity of 95.8% and a specificity of 63.2% in the pre-deep learning era [40]. Singapore had the first nationwide screening program launched in the 1990s that uses fundus images. There were later attempts to use the fundus images for developing AI models that can detect multiple eye diseases like diabetic retinopathy, glaucoma, and age-related macular-edema (AMD) [60]. They worked on the DR referability problem and achieved 93.6%, 90.5%, and 91.6% for the AUC, sensitivity, and specificity respectively.

They also reached 95.8%, 100%, and 91.1% for vision-threatening problem's AUC, sensitivity, and specificity respectively. Recently, Thailand also launched its DR screening program [46].Inspired by these attempts, we developed our own system.

3. DR10K Dataset

The first phase is to collect fundus images from diabetes patients. For each patient's eye, we collect both the macula-centered and the optic-disc-centered fundus images. This dataset (macula-centered) is then graded by ophthalmologists with PIRC scale grades for Diabetic Retinopathy and for DME existence. More details about the data collection and annotation system can be found in [4]

3.1. Data Distributions

We will show the dataset distribution with respect to gradability, DR levels, and referable DME.

3.1.1 Gradability

We collected 11,109 exams. The exams consist of either left or right fundus images of the patient or both. 87% of the images are gradable for both left and right eyes. Ungradable left-eye images in the dataset represent less than 5% of the total left-eye images. The same percentage holds for the right images. Also, we notice that 1825 exams have the image of only one eye and the other eye's image could not be captured which represents 16.4% of the whole exams. Finally, in total, we have 19,378 gradable images and 1,009 ungradable images. Our analysis shows also that only 10,811 exams have at least one gradable eye image. So, the other 298 exams are excluded from the rest of the analysis.

3.1.2 Diabetic Retinopathy Levels (DR-Levels)

For the remaining 19,378 images of 10,811 patients, we show the distribution of the DR levels in figure 2 and we show examples of our dataset in figure 1.

3.1.3 Diabetic Macular Edema (DME-Referability)

For the remaining 19,378 images of 10,811 patients, we show the distribution of the DME referability condition in table 1.

	Eyes Number	Patient Worst Eye
Non-referable DME	16397 (84.62 %)	8654 (80.05 %)
Referable DME	2981 (15.38 %)	2157 (19.95 %)

Table 1. Left column shows the number of eyes (right + left) graded as having DME or not. Right column shows the number of patients whose worst eye was graded as having DME or not.

3.2. Grader Variability

For each image, there were three graders and an adjudicator for the conflicts. In this section, we study the variability of the ophthalmologists' grades for the collected data. The aim of this study is to establish confidence in the grading procedure described in [4]. It is also to show the human grader baseline for the different classification tasks that are addressed. Establishing human-level metrics is a common practice while providing new datasets. The human performance was measured on medical datasets such as CheXpert [22] and on DR and DME datasets as in [59] and [61].

We started with 19,378 gradable images. Out of these images, we found that the three graders agreed on the gradability of 18,248 of them. In 995 images there were two graders in agreement and the adjudicator agreed with them. In 135 images there were two graders in agreement but the adjudicator disagreed with them (agreed with the third grader).

In the following, we will study the variability in the DR level with respect to the 18,248 images that were agreed on their gradability earlier. Table 2 shows that only 71.83% of the images did not need any adjudication. This observation confirms that depending on one ophthalmologist is not sufficient for DR classification and dataset annotation. In table 3, we show the graders' variability in the two binary problems of Referable DR and DR vision-threatening.

We establish the human grader baselines based on two approaches. In both approaches, the ground truth is considered an adjudicated grade. In the first approach, we contrast the ground truth against individual grading by the graders. In the second approach, we contrast the majority voting of the three graders to the ground truth. In table 4 we observe that the majority voting of the ophthalmologists produced much higher metrics. This observation hints to us to apply ensemble methods in our models that will be presented in 5.

4. Datasets and Splits Distributions

Training deep learning models requires a large amount of data. Usually, the performance of the model improves as the data amount increases. As the number of images in our new dataset is 19,378, which is relatively small, the model's training will suffer. So, inspired by previous work [17, 44] we used two publicly available datasets. We provide a table showing a comprehensive comparison between all the used datasets and a figure showing the DR levels distribution of the two public datasets in [5].

4.1. Kaggle Dataset

We downloaded a very large dataset from the Kaggle website. It is a combination of the data provided in two competitions, the first was held in 2015 during which the data is provided by EyePacs. The second is held in 2019

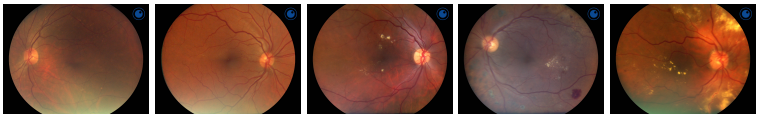

Figure 1. Examples of DR10K starting from grade 0 on the left to grade 4 on the right.

	3 Agreed(R)	2 Agreed (R) 1 Disag. (W)	2 Agreed (W) 1 Disag. (R)	2 Agreed (W) 1 Disag. (W)	3 Disag. 1 (R)	3 Disag. 3 (W)	Total
R0	11372	1402	650	2	8	0	13434 (73.62 %)
R1	1054	792	492	20	13	0	2371 (12.99 %)
R2	164	426	340	25	22	0	977 (5.35 %)
R3	160	378	150	14	18	0	720 (3.95 %)
R4	358	244	111	17	16	0	746 (4.09 %)
Total	13108 (71.83 %)	3242 (17.77 %)	1743 (9.55 %)	78 (0.43 %)	77 (0.42 %)	0 (0 %)	18248

Table 2. Graders' agreement percentages for DR levels.

	3 agreed (R)	3 agreed (W)	2 agreed (R) 1 disagreed (W)	2 agreed (W) 1 disagreed (R)
NRDR	15259	9	321	216
RDR	1653	1	534	255
Total	16912	10	855	471
Percent	92.68 %	0.05 %	4.69 %	2.58 %
NVT	16344	4	221	213
VT	802	15	485	164
Total	17146	19	706	377
Percent	93.96 %	0.1 %	3.87 %	2.07 %

Table 3. Grader's agreement percentages for binary DR problems.

	Individual Grading	Majority Voting
Human Accuracy	87 %	89.87 %
Human QWK	90.25 %	92.95 %
Human Binary Referability Accuracy	96.66 %	97.36 %
Human Binary Referability Specificity	98.35 %	98.58 %
Human Binary Referability Sensitivity	85.71 %	89.48 %
Human Binary VT Accuracy	97.23 %	97.82 %
Human Binary VT Specificity	98.69 %	98.66 %
Human Binary VT Sensitivity	80.49 %	88.27 %

Table 4. Human graders annotation performance metrics.

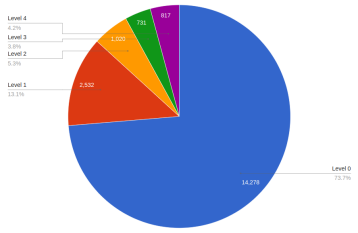

Figure 2. Per level DR10K dataset distribution

	Train	Validation	Test
Non ME	11715 (84.37%)	1552 (84.72%)	3130 (85.52%)
ME	2171 (15.63%)	280 (15.28%)	530 (14.48%)
Total	13886	1832	3660

Table 5. DME referabilty on DR10K dataset splits distribution

in which the data is provided by Aptos. The whole dataset is composed of 92,363 retinal images each labeled for DR severity from 0 to 4 according to the PIRC scale. We split the Kaggle data into three splits train, validation, and test. The test split is the public test set of the 2015 EyePacs competition. The train and validation splits are randomly sampled from the mix of the rest of the Kaggle data so as to keep the distribution of each of them similar to that of the whole dataset.

4.2. Messidor-2

We also used the publicly available Messidor-2 dataset [9] and [1] as additional testing data to prove the generalization of our model's performance.

4.3. DR10K

Finally, our dataset is split randomly into train, validation, and test splits. The splitting is done to keep the distribution of each split similar to the original distribution of the dataset shown in figure 2. Also, during splitting each patient's two eye images are included in the same split to keep the three splits independent. For DME existence classification, we used the same three splits that we acquired when splitting based on the DR distribution. In table 5, we show the distribution of the splits with respect to DME existence.

5. Approach

Our training process is supposed to train DR classification models using the larger Kaggle dataset then test it on Messidor-2 to prove generalization and finally fine-tune it on our Egyptian dataset to perform better on its scope. We will also fine-tune DR Kaggle-trained models on the DME labeled data to classify the DME existence.

5.1. Transfer Learning

Inception-V3 [57], DenseNet [19], MobileNetV2 [52], ResNet [18], ViT [12], Swin [35], ConvNeXt [36] and EfficientNet [58] are different deep architectures that achieved state-of-the-art results on ImageNet. Each one of our models is composed of one of these architectures pre-trained on ImageNet as a backbone with its last output layer replaced by one of our two classification blocks.

For the DR grading problem, we tried two options for the classification block the big classifier block and the small classifier block. The big classifier block is composed of four consecutive fully connected linear layers of sizes 1000, 500, 200, and 5 respectively. The small classifier block is composed of 3 consecutive fully connected linear layers of sizes 768, 256, and 5 respectively. We used the ReLU nonlinearity layer after each one of these linear layers except after the last layer where we used softmax. Moreover, we applied dropout layers with a probability of 0.5 between each pair of these linear layers during training. The complete architecture is shown in figure 3.

For the DME existence classification models, the last linear layer of the big classification block is replaced by a linear layer of size 1 instead of 5. Also, sigmoid nonlinearity is applied after this layer instead of softmax.

5.1.1 Implementation Details

We list the important details of our training procedures to ensure the reproducibility of the presented results.

- Any black borders or camera watermarks are removed from the fundus image before using it at any stage.
- We used Adam optimizer with weight decay of 4e-5, batch size of 32, 4 V100 GPUs, and tried the learning rates 5e-4, 1e-4, 5e-5, and 1e-5 choosing between them based on validation performance.
- The loss function is the cross entropy for DR models and the binary cross-entropy for DME existence ones.
- The images are resized to 880x880 then we apply augmentation techniques of random-resized-crop to a size of 768x768 (these sizes are fixed in all training phases except for ViT and Swin models where we resize to 440x440 then crop to 384x384). Horizontal and vertical flipping are also applied with a probability of 0.5.
- During the validation and testing stages the image is resized to 880x880 (440 for ViT and Swin) and a size of 768x768 (384 for ViT and Swin) is center cropped.
- During the 3 stages, ImageNet normalization is applied to the image.
- We trained our models for 100 epochs and chose the best-performing epoch on the validation set to be our experiment output model.

5.1.2 Baseline Models

To apply transfer learning to our classification problems, we employed several baseline models.

- EfficientNet-b7 and EfficientNet-b5 models pre-trained on ImageNet with the advprop [58] enabled.
- Inception-v3 model with auxiliary logits [57] disabled.
- Densenet-161 model.
- MobileNet-V2 model.
- ResNet152 model.
- ViT model with image size 384.
- Swin Transformer model with image size 384.
- ConvNeXt model.
- Ensemble of three different Architectures.
- Ensemble of three EfficientNet models.

5.1.3 Regression Score Binary Thresholding

Converting the classification probabilities to a regression score can be done easily in problems where the classes are meaningful levels such as five-class DR classification using the following equation 1

$$S = \Sigma_{c=0}^{4}\, c \cdot p_c \tag{1}$$

Where S is the regression score, c is the DR class ranges from 0 to 4 and p_c is the classifier's estimated probability of class c [26]. This score ranges from 0 to 4 and we can use it to classify referability and vision-threatening by looking for the best threshold over the validation set. We will choose a referability threshold and a vision-threatening threshold for each model and ensemble.

The choice of threshold introduces a trade-off between specificity and sensitivity. Since its value is directly proportional to the specificity and inversely proportional to the sensitivity. To guarantee the balance between them, we chose the threshold that achieves the highest harmonic mean of both of them on the validation data. We tried all the thresholds ranging from 0.05 to 3.95 with step 0.05. The chosen threshold is then tested on the appropriate test splits.

5.2. Augmentation using Weakly Labeled Data

After producing the results using our first approach, we looked for a way to get use of the optic-disc-centered images. The intuition behind that is to introduce a new view of the fundus data so that the model can learn to capture better features that aid it in the classification task. The optic-disc images can happen to have better quality or have some more clear features (lesions) that relate to the grade given to the eye.

The way we chose to integrate the optic-disc-centered in the training is by transferring the labels of the macula images to their corresponding optic-disc-centered images. In this case, we have a new weakly labeled dataset. This data is then added to the macula data to produce a new one of double the size (macula and optic-disc dataset). The label transfer procedure occurred in both DR and DME problems.

This dataset is then used in fine-tuning. Fine-tuning starts with the kaggle-trained checkpoints. We only investigated the best models according to the baseline results, hence we only fine-tuned the EfficientNet architecture on both DR and DME problems. The same preprocessing steps as well as the regression score thresholding technique were conducted.

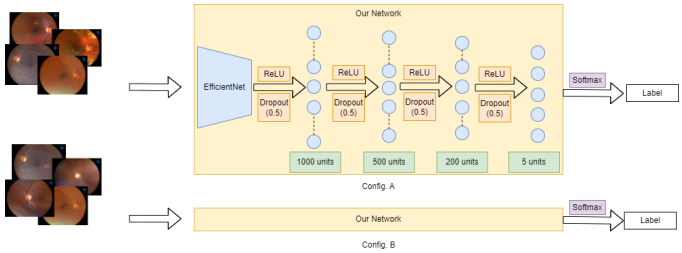

Figure 3. Our neural network architecture; In configuration A we only feed the macula-centered images to the network for fine-tuning, while in configuration B we feed the macula and the weakly labeled optic-disc centered images.

6. Results

6.1. Results on External Datasets

The baseline models are first trained on the Kaggle training dataset, and evaluated on the Kaggle test set and Messidor-2. The aim is to use the models with the best results for fine-tuning DR10K. We found that the ensemble of three EfficientNet models is doing the best. Also, the best standalone model has the EfficientNet backbone. Detailed results are shown in [5]. Hence, only EfficientNet models are fine-tuned.

6.2. Results on DR10K Dataset

6.2.1 Results for DR

In table 6 we compare three alternatives for the five-class grading problem. We examine the best models on the Kaggle dataset against the test set of the DR10K dataset. We also examine the option of training EfficientNet models from scratch on the training set of DR10K. We examine the option of fine-tuning the best models we obtained earlier on the training set of DR10K. The results show that the models that are trained from scratch on DR10K or fine-tuned performed much better than evaluating the models trained on Kaggle directly. Moreover, we sampled a subset from the DR10K test split similar in distribution to the Kaggle test split and tested the Kaggle-trained models on it. The results are better than direct testing on DR10K distribution but still worse than models trained from scratch or fine-tuned on DR10K. This proves that the data distribution is an important factor leading to performance degradation but is not the only one. This shows the importance of gathering Egyptian data for the screening program to reflect the local distribution, used fundus camera, and ethnic biological features. The importance of fine-tuning is also evident since the fine-tuned models on DR10K perform better than the models trained from scratch. Also, the best result we achieved exceeds the human baseline for individual grading. The effect of augmentation using weakly labeled data is also evident in the standalone and the ensemble of models.

	Accuracy	QWK
Kaggle Trained		
Eff-b5 -small+ lr 1e-4	79.78	78.89
Eff-b7 -small+ lr 5e-5	78.55	77.25
Eff-b7-big+ lr 1e-4	**81.12**	81.3
Ensemble of 3 EfficientNet models	80.85	**81.61**
Kaggle Trained Tested on Distribution Similar to Kaggle		
Eff-b5 -small+ lr 1e-4	84	82.64
Eff-b7 -small+ lr 5e-5	80.85	76.73
Eff-b7-big+ lr 1e-4	**86.07**	**83.74**
Trained From Scratch on DR10K		
Eff-b5 -small+ lr 1e-4	88.28	89.69
Eff-b7 -small+ lr 5e-5	88.47	89.42
Eff-b7-big+ lr 1e-4	87.76	88.66
Ensemble of 3 EfficientNet models	**88.99**	**90.06**
DR10K Fine Tuned		
Eff-b5 -small+ lr 1e-4	88.88	89.82
Eff-b7 -small+ lr 5e-5	89.23	89.91
Eff-b7-big+ lr 1e-4	89.21	89.7
Ensemble of 3 EfficientNet models	**89.51**	**90.23**
DR10K Fine Tuned with optic-disc		
Eff-b7 -big+ lr 1e-5	89.51	89.33
Eff-b7 -big+ lr 5e-5	88.83	90.09
Eff-b7 -big+ lr 5e-4	88.77	90.16
Ensemble of 3 EfficientNet models	**89.86**	**91.3**

Table 6. Five-Class DR Classification Performance on DR10K.

As for the binary problems and the use of the regression score thresholding technique, the results are shown in tables 7 and 8. We observe that the fine-tuned models on the DR10K training are performing much better than the other alternatives except for the models trained on the extended data with optic-disc images. One of these models has beaten all standalone macula fine-tuned models and has achieved a comparable performance to the macula ensemble for the referability problem. Also, the ensemble of models fine-tuned with the extended data outperforms the ensemble of fine-tuned models on the macula in most of the metrics for the vision-threatening problem.

6.2.2 Results for DME referability

We compare three scenarios for the DME binary problem. We examine training EfficientNet models from scratch (pre-

	Thr.	Acc.	Sens.	Spec.	H-Mean	AUC
			Kaggle Trained			
Eff-b5-small + lr 1e-4	0.7	**87.57**	96.72	**86.26**	**91.19**	96.75
Eff-b7-small + lr 5e-5	0.6	83.06	**98.91**	80.8	88.94	97.29
Eff-b7-big + lr 1e-4	0.55	85.96	98.91	84.11	90.91	**97.9**
Ensemble of 3 models	0.7	86.34	98.69	84.58	91.09	97.84
			DR10K Trained From Scratch			
Eff-b5-small + lr 1e-4	1.15	94.97	96.06	94.82	**95.44**	98.88
Eff-b7small + lr 5e-5	1	93.17	96.72	92.66	94.65	98.98
Eff-b7-big + lr 1e-4	1.05	**95.03**	94.97	**95.04**	95	98.76
Ensemble of 3 models	1.05	94.02	**97.16**	93.57	95.33	**99.12**
			DR10K Fine Tuned			
Eff-b5-small + lr 1e-4	1.05	95.25	**97.16**	94.97	96.05	99.16
Eff-b7-small + lr 5e-5	1.1	95.19	95.4	95.16	95.28	99.18
Eff-b7-big + lr 1e-4	1.1	94.32	96.5	94.01	95.24	99.17
Ensemble of 3 models	1.15	**95.79**	96.94	**95.63**	**96.28**	**99.41**
			DR10K Fine Tuned with optic-disc			
Eff-b7 -big+ lr 1e-5	1.1	94.26	97.16	93.85	95.47	99.22
Eff-b7 -big+ lr 5e-5	1.05	94.51	96.28	94.26	95.26	99.12
Eff-b7 -big+ lr 5e-4	1.15	**95.44**	97.16	**95.19**	**96.16**	99.28
Ensemble of 3 models	1.1	94.59	**97.59**	94.16	95.85	**99.36**

Table 7. DR Referability Classification Performance on DR10K.

	Thr.	Acc	Sens.	Spec.	H-Mean	AUC
			Kaggle Trained			
Eff-b5 small + lr 1e-4	2	91.91	85.46	**92.45**	88.82	95.94
Eff-b7 small + lr 5e-5	2.05	91.17	**92.2**	91.09	**91.64**	96.87
Eff-b7 big + lr 1e-4	1.9	91.17	91.49	91.15	91.32	97.12
Ensemble of 3 models	2	**92.19**	90.07	92.36	91.2	**97.22**
			DR10K Trained From Scratch			
Eff-b5 small + lr 1e-4	1.75	**94.62**	**95.74**	**94.52**	**95.13**	98.5
Eff-b7 small + lr 5e-5	1.45	94.02	94.33	93.99	94.16	98.4
Eff-b7 big + lr 1e-4	1.7	94.13	93.26	94.2	93.73	98.34
Ensemble of 3 models	1.65	94.48	94.33	94.49	94.41	**98.69**
			DR10K Fine Tuned			
Eff-b5 small + lr 1e-4	1.55	94.89	94.33	94.94	94.63	98.68
Eff-b7 small + lr 5e-5	1.6	95.03	93.26	95.17	94.21	98.6
Eff-b7 big + lr 1e-4	1.9	**95.96**	**96.35**	**95.93**	**96.14**	98.46
Ensemble of 3 models	1.55	94.64	95.39	94.58	94.98	**98.87**
			DR10K Fine Tuned with optic-disc			
Eff-b7 -big+ lr 1e-5	1.75	**95.38**	91.13	**95.74**	93.38	98.73
Eff-b7 -big+ lr 5e-5	1.55	94.37	93.97	94.4	94.19	98.75
Eff-b7 -big+ lr 5e-4	1.6	94.56	**97.16**	94.35	**95.73**	98.79
Ensemble of 3 models	1.5	93.83	96.45	93.61	95.01	**98.92**

Table 8. DR Vision-Threatening Classification Performance on DR10K.

trained only on ImageNet) using DR10K. We also examine the transfer learning from the DR problem to the DME, we assumed the similarity between features can help in the DME training. We used the models trained on the Kaggle Dataset for DR as the starting models for the fine-tuning step. We also used the optic-disc extended data in the DME fine-tuning. We can find in tables 9 and 10 that the results are comparable yet the models trained from scratch performed slightly better.

We can notice that regression score thresholding has a great impact on preserving the balance between sensitivity and specificity for this problem. So, we report the results of the extended dataset following this technique and we notice that ensemble accuracy outperforms other techniques.

	Acc.	Sens.	Spec.	H-Mean	AUC
		DR10K Trained From Scratch			
Eff-b5-small + lr 1e-4	94.43	**91.89**	94.86	**93.35**	98.58
Eff-b7-small + lr 5e-5	95.41	85.66	**97.06**	91	98.57
Eff-b7-big + lr 1e-4	95.46	87.92	96.74	92.12	95.46
Ensemble of 3 models	**95.74**	90.19	96.68	93.32	**98.83**
		DR10K Fine Tuned			
Eff-b5-small + lr 1e-4	94.75	**88.11**	95.88	91.83	98.31
Eff-b7-small + lr 5e-5	**95.44**	85.66	**97.09**	91.02	98.55
Eff-b7-big + lr 1e-4	95	87.55	96.26	91.7	98.65
Ensemble of 3 models	95.33	87.55	96.65	**91.87**	**98.77**

Table 9. DME Existence Classification Performance on DR10K.

	Thr.	Acc.	Sens.	Spec.	H-Mean
			DR10K Trained From Scratch		
Eff-b5-small + lr 1e-4	0.3	92.9	95.28	92.49	93.86
Eff-b7-small + lr 5e-5	0.1	92.21	**96.79**	91.44	94.03
Eff-b7-big + lr 1e-4	0.1	94.43	93.4	**94.6**	93.99
Ensemble of 3 models	0.3	**94.48**	94.34	94.5	**94.41**
			DR10K Fine Tuned		
Eff-b5-small + lr 1e-4	0.25	92.95	95.47	92.52	93.97
Eff-b7-small + lr 5e-5	0.05	93.42	96.04	92.97	94.48
Eff-b7-big + lr 1e-4	0.1	92.76	**97.36**	91.98	94.59
Ensemble of 3 models	0.25	**94.67**	95.85	**94.47**	**95.15**
			DR10K Fine Tuned with optic-disc		
Eff-b7 -big+ lr 1e-4	0.251	92.9	95.09	92.52	93.78
Eff-b7 -big+ lr 5e-5	0.151	93.52	**95.47**	93.19	**94.31**
Eff-b7 -big+ lr 5e-4	0.201	93.63	92.83	93.77	93.29
Ensemble of 3 models	0.351	**94.92**	91.7	**95.46**	93.54

Table 10. DME Existence Classification Performance using Regression Score Thresholding on DR10K.

7. Discussion

We will discuss some of the findings that we found to be crucial through our experiments. Then, we will highlight our proposed DR and DME screening pipeline steps.

7.1. Findings

- When using deep learning models, transfer learning is crucial to compensate for the fact that the data collection process is expensive. For certain problems like ours, fine-tuning on local data is also essential. This was shown when we tested Kaggle-trained models on our Egyptian data and the performance degraded drastically 6. This is quantitative proof for the assumptions made by previous work in [49] that the DR classification performance degrades when the used models are trained on different ethnicity data.

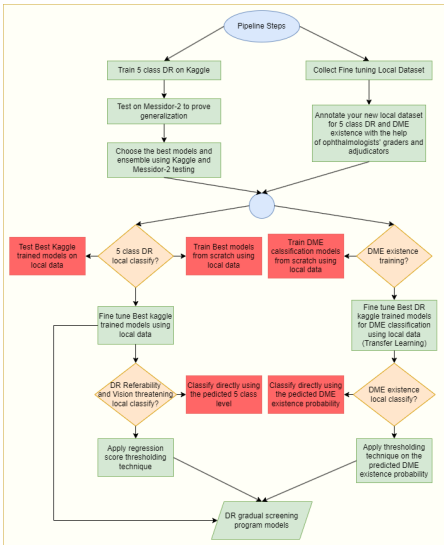

Figure 4. Pipeline for establishing a deep learning-based system for DR and DME screening programs.

- In the DME referability classification, the training from scratch using Egyptian data seems to work fine as shown in table 9. This is different from what happened on the DR classification task where fine-tuning Kaggle models worked better. We can observe that for the DME existence problem, the percentage of the minor positive class of the dataset is more than 15%. On the other hand, the minor classes 3 and 4 of the DR classification problem represent only 3.77% and 4.22% of the whole data respectively. We hypothesize that this difference in the class imbalance degree may be the main reason that leads to the performance difference when training from scratch in each problem.
  The results after applying regression score thresholding shown in table 10, show that the fine-tuned models slightly outperformed the models trained from scratch. This means that the transfer learning process from DR to DME succeeds and helps in enhancing the DME existence classification performance.
- It is crucial to achieve high sensitivity and harmonic mean in our binary classification problems. The importance of these binary problems arises from the fact that they determine the system decision for the patient. Since the collection of more positive data is time-consuming process, we applied the regression score thresholding technique. This technique improved the sensitivity significantly while preserving high specificity. This effect is shown in tables 7, 8, and 10.

7.2. Pipeline

Based on our observations we propose a complete pipeline for training the models needed in the DR screening program. The pipeline steps are as follows:

1. Train DR classification models using Kaggle dataset.
2. Test the Kaggle-trained models using the Messidor-2 dataset as a generalization proof.
3. Collect a local small finetuning dataset including both macula and optic-disk-centered images for each eye.
4. Annotate the local dataset for both DR levels and DME existence using the macula-centered images.
5. Propagate the labels of macula-centered images to the corresponding optic-disk-centered ones.
6. Fine-tune Kaggle-trained models using the full local dataset for DR classification.
7. Apply transfer learning technique by fine-tuning DR Kaggle-trained models for DME existence classification using the full local dataset.
8. Apply binary thresholding technique to enhance the performance in the three binary problems (referability, vision-threatening, and DME referabiltiy).

Figure 4 shows all our experimented alternatives highlighting in green the best-chosen pipeline steps that achieved our state-of-the-art results.

8. Conclusion

The paper proposes a complete pipeline that enables training the models needed for the DR regular screening program to achieve state-of-the-art results taking into consideration both the time and cost factors. We succeeded in achieving a good performance using the publicly available Kaggle dataset. The Kaggle-trained models are proved to be generalizable achieving good results on the Messidor-2 dataset reaching a 5 class accuracy of 83.6% and QWK of 90.06%. We collected our new Egyptian dataset and annotated it for both DR levels and DME referability. We provide a full analysis of this dataset showing all their insights and highlighting the human ophthalmologists' classification performance in the graders' variability subsection.

We compared different training methods and discussed the reasons behind their performance. We enhanced the training with the extended dataset of optic-disc centered images and it succeded in outperforming our previous methods in DR. It achieved a comparable performance in the DME problem. Our introduced complete pipeline can represent a road map for any developing country.

9. Acknowledgement

The authors would like to thank the Applied Innovation Center (AIC) of Egyptian Ministry of Communication and Information Technology for funding the research presented in this paper. The authors would like to thank Dr. Ahmed Tantawy the director of AIC for initiating this project.

References

[1] Michael D Abràmoff, James C Folk, Dennis P Han, Jonathan D Walker, David F Williams, Stephen R Russell, Pascale Massin, Beatrice Cochener, Philippe Gain, Li Tang, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA ophthalmology*, 131(3):351–357, 2013. 4

[2] M Usman Akram, Shehzad Khalid, and Shoab A Khan. Identification and classification of microaneurysms for early detection of diabetic retinopathy. *Pattern recognition*, 46(1):107–116, 2013. 1

[3] Wejdan L Alyoubi, Wafaa M Shalash, and Maysoon F Abulkhair. Diabetic retinopathy detection through deep learning techniques: A review. *Informatics in Medicine Unlocked*, 20:100377, 2020. 1

[4] Authors. Appendix a. dr10k data collection system details, 2023. Supplied as supplemental material `Appendix A.pdf`. 3

[5] Authors. Appendix b. miscellaneous, 2023. Supplied as supplemental material `Appendix B.pdf`. 3, 6

[6] Neelakshi Bhagat, Ruben A Grigorian, Arthur Tutela, and Marco A Zarbin. Diabetic macular edema: pathogenesis and treatment. *Survey of ophthalmology*, 54(1):1–32, 2009. 1

[7] Titus Josef Brinker, Achim Hekler, Jochen Sven Utikal, Niels Grabe, Dirk Schadendorf, Joachim Klode, Carola Berking, Theresa Steeb, Alexander H Enk, and Christof Von Kalle. Skin cancer classification using convolutional neural networks: systematic review. *Journal of medical Internet research*, 20(10):e11936, 2018. 2

[8] Genevieve C. Y. Chan, Ravi Kamble, Henning Muller, Syed A. A. Shah, T. B. Tang, and Fabrice Meriaudeau. Fusing results of several deep learning architectures for automatic classification of normal and diabetic macular edema in optical coherence tomography. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, July 2018. 2

[9] Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, 33(3):231–234, 2014. 4

[10] S Deepak and PM Ameer. Brain tumor classification using deep cnn features via transfer learning. *Computers in biology and medicine*, 111:103345, 2019. 2

[11] Anamika Dhillon and Gyanendra K Verma. Convolutional neural network: a review of models, methodologies and applications to object detection. *Progress in Artificial Intelligence*, 9(2):85–112, 2020. 2

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[13] Ismail El-Yamany, Abdelrahman Wael, Noha Adly, and Marwan Torki. Stackmaps: A visualization technique for diabetic retinopathy grading. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2

[14] Ahmed Elmassry, Islam SH Ahmed, Noha Adly, and Marwan Torki. Prevalence of diabetic retinopathy in patients with diabetes in alexandria and north-west delta, egypt. *International Ophthalmology*, pages 1–13, 2023. 2

[15] Oliver Faust, Rajendra Acharya U, Eddie Yin-Kwee Ng, Kwan-Hoong Ng, Jasjit S Suri, et al. Algorithms for the automated detection of diabetic retinopathy using digital fundus images: a review. *Journal of medical systems*, 36(1):145–157, 2012. 1

[16] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Multimodal semi-supervised learning for image classification. In *2010 IEEE Computer society conference on computer vision and pattern recognition*, pages 902–909. IEEE, 2010. 2

[17] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016. 3

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[19] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 4

[20] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1055–1059. IEEE, 2020. 2

[21] Tri Huynh, Aiden Nibali, and Zhen He. Semi-supervised learning for medical image classification using imbalanced training data. *Computer Methods and Programs in Biomedicine*, page 106628, 2022. 2

[22] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019. 3

[23] Zhipeng Jia, Xingyi Huang, I Eric, Chao Chang, and Yan Xu. Constrained deep weak supervision for histopathology image segmentation. *IEEE transactions on medical imaging*, 36(11):2376–2388, 2017. 2

[24] Zhencun Jiang, Lingyang Wang, Qixin Wu, Yilei Shao, Meixiao Shen, Wenping Jiang, and Cuixia Dai. Computer-aided diagnosis of retinopathy based on vision transformer. *Journal of Innovative Optical Health Sciences*, 15(02):2250009, 2022. 2

[25] Sharif Amit Kamran, Khondker Fariha Hossain, Alireza Tavakkoli, Stewart Lee Zuckerbrod, and Salah A Baker. Vtgan: Semi-supervised retinal image synthesis and disease prediction using vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3235–3245, 2021. 2

[26] Jonathan Krause, Varun Gulshan, Ehsan Rahimy, Peter Karth, Kasumi Widner, Greg S Corrado, Lily Peng, and Dale R Webster. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*, 125(8):1264–1272, 2018. 5

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 2

[28] Carson Lam, Darvin Yi, Margaret Guo, and Tony Lindsey. Automated detection of diabetic retinopathy using deep learning. *AMIA summits on translational science proceedings*, 2018:147, 2018. 1

[29] Janet L Leasher, Rupert RA Bourne, Seth R Flaxman, Jost B Jonas, Jill Keeffe, Kovin Naidoo, Konrad Pesudovs, Holly Price, Richard A White, Tien Y Wong, et al. Global estimates on the number of people blind or visually impaired by diabetic retinopathy: a meta-analysis from 1990 to 2010. *Diabetes care*, 39(9):1643–1649, 2016. 1

[30] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995. 2

[31] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 2

[32] Yann LeCun, J Denker, D Henderson, R Howard, W Hubbard, and L Jacke. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 1990. 2

[33] Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 34–42, 2015. 2

[34] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE transactions on medical imaging*, 37(12):2663–2674, 2018. 2

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 4

[36] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 4

[37] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *Asian Conference on Computer Vision*, pages 565–580. Springer, 2018. 2

[38] Ishan Misra, Abhinav Shrivastava, and Martial Hebert. Watch and learn: Semi-supervised learning for object detectors from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3593–3602, 2015. 2

[39] Ghulam Murtaza, Liyana Shuib, Ainuddin Wahid Abdul Wahab, Ghulam Mujtaba, Henry Friday Nweke, Mohammed Ali Al-garadi, Fariha Zulfiqar, Ghulam Raza, and Nor Aniza Azmi. Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges. *Artificial Intelligence Review*, 53(3):1655–1720, 2020. 2

[40] Carlos M Oliveira, Luis M Cristovao, Maria Luisa Ribeiro, and José R Faria Abreu. Improved automated screening of diabetic retinopathy. *Ophthalmologica*, 226(4):191–197, 2011. 2

[41] Myeongsuk Pak and Sanghoon Kim. A review of deep learning in image recognition. In *2017 4th international conference on computer applications and information processing technology (CAIPT)*, pages 1–3. IEEE, 2017. 2

[42] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1742–1750, 2015. 2

[43] Grzybowski Pieczynski J. A. review of diabetic retinopathy screening methods and programmes adopted in different parts of the world. *European Ophthalmic Review*, 9:49– 55, 2015. 2

[44] Harry Pratt, Frans Coenen, Deborah M Broadbent, Simon P Harding, and Yalin Zheng. Convolutional neural networks for diabetic retinopathy. *Procedia computer science*, 90:200–205, 2016. 3

[45] Holger R Roth, Dong Yang, Ziyue Xu, Xiaosong Wang, and Daguang Xu. Going to extremes: weakly supervised medical image segmentation. *Machine Learning and Knowledge Extraction*, 3(2):507–524, 2021. 2

[46] Paisan Ruamviboonsuk, Richa Tiwari, Rory Sayres, Variya Nganthavee, Kornwipa Hemarat, Apinpat Kongprayoon, Rajiv Raman, Brian Levinstein, Yun Liu, Mike Schaekermann, et al. Real-time diabetic retinopathy screening by deep learning in a multisite national screening programme: a prospective interventional cohort study. *The Lancet Digital Health*, 4(4):e235–e244, 2022. 3

[47] Pouya Saeedi, Inga Petersohn, Paraskevi Salpea, Belma Malanda, Suvi Karuranga, Nigel Unwin, Stephen Colagiuri, Leonor Guariguata, Ayesha A Motala, Katherine Ogurtsova, et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the international diabetes federation diabetes atlas. *Diabetes research and clinical practice*, 157:107843, 2019. 1

[48] Hamid Safi, Sare Safi, Ali Hafezi-Moghadam, and Hamid Ahmadieh. Early detection of diabetic retinopathy. *Survey of ophthalmology*, 63(5):601–608, 2018. 1

[49] Jaakko Sahlsten, Joel Jaskari, Jyri Kivinen, Lauri Turunen, Esa Jaanio, Kustaa Hietala, and Kimmo Kaski. Deep learning fundus image analysis for diabetic retinopathy and mac-

ular edema grading. *Scientific reports*, 9(1):1–11, 2019. 1, 2, 7

[50] Jaakko Sahlsten, Joel Jaskari, Jyri Kivinen, Lauri Turunen, Esa Jaanio, Kustaa Hietala, and Kimmo Kaski. Deep learning fundus image analysis for diabetic retinopathy and macular edema grading. *Scientific reports*, 9(1):1–11, 2019. 2

[51] Ahmed Salama, Noha Adly, and Marwan Torki. Ablation-cam++: Grouped recursive visual explanations for deep convolutional networks. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2011–2015. IEEE, 2022. 2

[52] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 4

[53] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[54] Chanjira Sinthanayothin, James F Boyce, Tom H Williamson, Helen L Cook, Evelyn Mensah, Shantanu Lal, and David Usher. Automated detection of diabetic retinopathy on digital fundus images. *Diabetic medicine*, 19(2):105–112, 2002. 1

[55] Rui Sun, Yihao Li, Tianzhu Zhang, Zhendong Mao, Feng Wu, and Yongdong Zhang. Lesion-aware transformers for diabetic retinopathy grading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10938–10947, 2021. 2

[56] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 2

[57] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 2, 4, 5

[58] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 4, 5

[59] Daniel Shu Wei Ting, Carol Yim-Lui Cheung, Gilbert Lim, Gavin Siew Wei Tan, Nguyen D Quang, Alfred Gan, Haslina Hamzah, Renata Garcia-Franco, Ian Yew San Yeo, Shu Yen Lee, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *Jama*, 318(22):2211–2223, 2017. 1, 3

[60] Daniel Shu Wei Ting, Carol Yim-Lui Cheung, Gilbert Lim, Gavin Siew Wei Tan, Nguyen D. Quang, Alfred Gan, Haslina Hamzah, Renata Garcia-Franco, Ian Yew San Yeo, Shu Yen Lee, Edmund Yick Mun Wong, Charumathi Sabanayagam, Mani Baskaran, Farah Ibrahim, Ngiap Chuan Tan, Eric A. Finkelstein, Ecosse L. Lamoureux, Ian Y. Wong, Neil M. Bressler, Sobha Sivaprasad, Rohit Varma, Jost B. Jonas, Ming Guang He, Ching-Yu Cheng, Gemmy Chui Ming Cheung, Tin Aung, Wynne Hsu, Mong Li Lee, and Tien Yin Wong. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*, 318(22):2211, Dec. 2017. 2

[61] Avinash V. Varadarajan, Pinal Bavishi, Paisan Ruamviboonsuk, Peranut Chotcomwongse, Subhashini Venugopalan, Arunachalam Narayanaswamy, Jorge Cuadros, Kuniyoshi Kanai, George Bresnick, Mongkol Tadarati, Sukhum Silpaarcha, Jirawut Limwattanayingyong, Variya Nganthavee, Joseph R. Ledsam, Pearse A. Keane, Greg S. Corrado, Lily Peng, and Dale R. Webster. Predicting optical coherence tomography-derived diabetic macular edema grades from fundus photographs using deep learning. *Nature Communications*, 11(1), Jan. 2020. 2, 3

[62] Mike Voets, Kajsa Møllersen, and Lars Ailo Bongo. Reproduction study using public data of: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *PLOS ONE*, 14(6):e0217541, June 2019. 2

[63] Zhiguang Wang and Jianbo Yang. Diabetic retinopathy detection via deep convolutional networks for discriminative localization and visual explanation. *arXiv.org*, 2017. 2

[64] Moustafa Wassel, Ahmed M Hamdi, Noha Adly, and Marwan Torki. Vision transformers based classification for glaucomatous eye condition. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 5082–5088. IEEE, 2022. 2

[65] Yu Weng, Tianbao Zhou, Yujie Li, and Xiaoyu Qiu. Nas-unet: Neural architecture search for medical image segmentation. *IEEE Access*, 7:44247–44257, 2019. 2

[66] Tien Yin Wong and Neil M Bressler. Artificial intelligence with deep learning technology looks into diabetic retinopathy screening. *Jama*, 316(22):2366–2367, 2016. 1

[67] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019. 2

[68] Jiabin Zhang, Hu Su, Wei Zou, Xinyi Gong, Zhengtao Zhang, and Fei Shen. Cadn: a weakly supervised learning-based category-aware object detection network for surface defect detection. *Pattern Recognition*, 109:107571, 2021. 2

[69] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018. 2