**IDEATION**

**Original fundus image dataset**

* The research of present study was done in collaboration with Digifundus Ltd, an ISO 9001:2015 certified provider of diabetic retinopathy screening and monitoring services in Finland.
* Digifundus Ltd provided a non-open, anonymized retinal image dataset of patients with diabetes, including 41122 graded retinal color images from 14624 patients.
* The images were taken with Canon CR2 retinal camera after inducing mydriasis with tropicamide 5 mg/ml eyedrops.
* Two 45 degree color fundus photographs, centered on fovea and optic disc were taken from the patient’s both eyes.
* The output images were of variable resolutions, ranging from 3888 × 2592 to 5184 × 3456 pixels.

**Retinal image grading systems and gradability**

* Each of the retinal images had been graded with respect to three different criteria, (i) diabetic retinopathy, (ii) macular edema, and (iii) gradability.
* Images are graded with the proposed international clinical diabetic retinopathy and macular edema disease severity scales[6](https://www.nature.com/articles/s41598-019-47181-w#ref-CR6), denoted later as PIRC and PIMEC, respectively.
* The image gradability is a two-stage system, which considers an image to be either gradable or not.
* All personnel participating in retinopathy assessment had over 10 years’ experience in diabetic retinopathy grading. Retinal images with no lesions or mild diabetic lesions were graded by an optometrist and an M.D. trained for retinopathy grading.
* All images with moderate or worse changes were graded by two ophthalmologist both with more than 10 years of experience in grading fundus images. If there was a disagreement in grading, such an image was not included in this study.

**Image preprocessing and dataset division**

* In the model training and subsequent primary validation, we used preprocessed versions of the original images.
* The preprocessing consisted of image cropping followed by resizing. Each image was cropped to a square shape which included the most tightly contained circular area of fundus.
* The procedure removed most of the black borders and all of the patient related annotations from the image data.
* Each of the cropped images were then resized to five different standard input image sizes of 256 × 256, 299 × 299, 512 × 512, 1024 × 1024, and 2095 × 2095 pixels.
* The largest image size was the smallest native resolution of the retinal cameras after the preprocessing steps.
* The obtained processed datasets were divided into three sets: *training*, *tuning*, and *primary validation* set.

### Deep learning model

* In order to distinguish features related to diabetic retinopathy and macular edema in the color images of patients’ fundi we chose to use a deep convolutional neural network.
* The neural networks are mathematical models, which consist of parameters used in specific calculations, such as convolutions and summations.
* Here the neural network can be constructed in such a way, that it receives an input which is used in calculating an output[12](https://www.nature.com/articles/s41598-019-47181-w#ref-CR12), such as class or grade of diabetic retinopathy.
* The parameters used in the calculation of the output can be modified in a data-driven manner, by minimizing the error between neural network produced detections of classes and the manual annotations.

### Model evaluation and comparison against previous works

* The present study was conducted by training a separate model for each of the five classification tasks and five input image sizes.
* To evaluate the performance of our model in binary classification tasks we use the receiver operating characteristic (ROC) curve from which we determine the area under the ROC curve (AUC) as well as accuracy, sensitivity, and specificity, while in the multi-class cases we use the area under macro average of ROC (macro-AUC) for each class calculated in one-vs-all manner, accuracy and quadratic-weighted kappa score.
* Also, we calculate the confusion matrices for the multi-class classification tasks. For each metric in the binary classification tasks, the exact 95% confidence interval (CI) was calculated using the Clopper-Pearson method, similar to that in Gulshan *et al*.[4](https://www.nature.com/articles/s41598-019-47181-w#ref-CR4) for comparison.