Rethinking Retinal Image Quality

Rethinking Retinal Image Quality: Treating Quality Threshold as a Tunable Hyperparameter*

Fabian SL Yii[1], Raman Dutt[1], Tom MacGillivray[1], Baljean Dhillon[1], Miguel Bernabeu[1], and Niall Strang[2]

[1] University of Edinburgh, Edinburgh, UK
fabian.yii@ed.ac.uk
[2] Glasgow Caledonian University, Glasgow, UK

Abstract. Assuming the robustness of a deep learning model to suboptimal images is a key consideration, we asked if there was any value in including training images of poor quality. In particular, should we treat the (quality) threshold at which a training image is either included or excluded as a tunable hyperparameter? To that end, we systematically examined the effect of including training images of varying quality on the test performance of a DL model in classifying the severity of diabetic retinopathy. We found that there was a unique combination of (categorical) quality labels or a *Goldilocks* (continuous) quality score that gave rise to optimal test performance on either high-quality or suboptimal images. The model trained exclusively on high-quality images yielded worse performance in all test scenarios than that trained on the optimally tuned training set which included images with some level of degradation.

Keywords: Image quality · Tunable hyperparameter · Deep learning.

1 Introduction

A common pre-processing step in deep learning (DL) applied to retinal image analysis is to exclude images of sub-optimal quality before training and testing a model for a given downstream task. For instance, Poplin et al. filtered out 12% of 96,082 UK Biobank (UKBB) retinal images of 'poor quality' for a downstream task of predicting different cardiovascular risk factors [14]. Likewise, 12% of UKBB retinal images of 'very poor quality' were excluded in another study aiming to predict refractive error [19]. Lin et al. removed 14,003 retinal images that were 'subjectively' deemed to be poor – or if the optic disc and fovea were not present simultaneously – from a total of 35,126 EyePacs images, with a view to training a model to detect referable diabetic retinopathy (DR) [11].

The tacit assumption of removing poor images in the application of DL is that only input images of *relatively high quality* are to be used when a model

* F. Yii and R. Dutt contributed equally to this work. F. Yii is supported by the Medical Research Council [grant number MR/N013166/1]

is deployed in the real world. Such a model will, conceivably, not generalise well to images with some degradation arising from, say, naturally occurring senile eye conditions like cataract, or sub-optimal patient positioning leading to non-uniform illumination. We are therefore drawn to think that *careful* inclusion of images with some appropriate level of degradation may in fact make a model more *versatile*, i.e. robust to a wider distribution of image quality. Indeed, segmentation of retinal sublayers and choroid in optical coherence tomography (OCT) images improves when a DL model is trained on degraded images [9]. Similar observation has been made when classifying non-medical images with DL [3].

But how do we determine if a given training image has an appropriate level of degradation, such that it is high enough to add some useful noise but low enough to not undermine the model? We propose that the image quality threshold, at which we decide if an image should or should not be used for training, can be treated as a *tunable hyperparameter*. Our ultimate goal is to maximise the performance of a trained model on the *unfiltered* test set to simulate real-world distribution of image quality. This is in contrast to studies where the model is trained and tested on *filtered* datasets. While some may argue that a simpler approach is to *apply* various levels and types of image degradation [4], and settle on the level that yields the best test performance, we are of the opinion that such artificial image degradation, e.g. gaussian blur, is not nuanced enough to capture the kind of degradation particular to a retinal image, e.g. areas of under- and over-exposure during acquisition.

The idea that image quality threshold can be treated as a tunable hyperparameter raises the question of whether it should be done on a *categorical* or *continuous* scale. In this regard, it is conceivable that superior outcome (as judged by the test performance of a downstream model trained on the resultant, filtered dataset) is contingent upon one's ability to partition training images based on their quality at as *granular* a level as possible – since this renders any effect of nuanced variation in image quality discernible. Thus, the main objectives of our work are to see if:

- including images of poorer quality in the training set has a positive bearing on the test performance (particularly on the *unfiltered* test set) of a DL model for a downstream task of classifying DR severity. If so, should we treat quality threshold as a tunable hyperparameter?
- tuning the quality threshold on a *continuous* scale offers additional value (more optimal test performance) than tuning on a categorical scale. We hypothesise an increase in test performance (model made more robust) as the training set becomes noisier up to a point before falling.

2 Methods

2.1 Quality prediction on a categorical scale and continuous scale

To predict image quality on a categorical scale, we utilised *Multiple Color-space Fusion Network* (MCF-Net), a DL model that achieved a state-of-the-art test

accuracy of 91.75% [6]. Briefly, an image is considered *good* if there are no low-quality factors; *usable* if there are some low-quality factors but important features like the optic disc are still clear enough for ophthalmological assessment to be carried out; *reject* if a full assessment is impossible.

Description of the adapted (regression) model. To turn the original model into a regression model, we removed the softmax function corresponding to each of the 5 loss functions. Mean absolute error (MAE) was used in place of the original cross-entropy loss function. The output of the adapted model (normalised between 0 and 1) would be *closer to 0 for a high-quality image*. The model achieved an MAE of 0.154 on the test set. More information on the adapted model is available as supplementary material (S1).

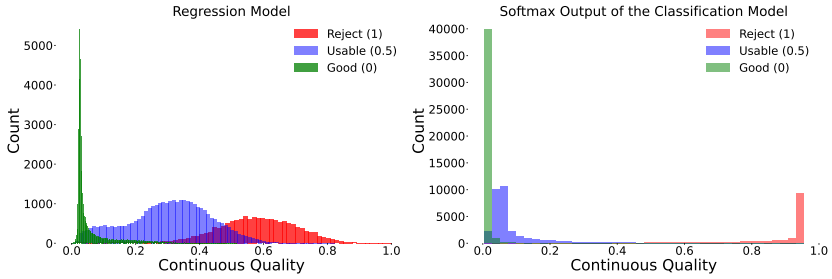

Fig. 1. Distribution of continuous quality scores of the entire *EyePACS* dataset (n=88,702) as predicted by the regression model (left) and as represented by the softmax output (*Reject* class, i.e. greater value corresponds to poorer quality) of the original classification model (right). Each hue represents a different quality class.

As a baseline approach (owing to its simplicity), we also extracted the softmax output in the final classification layer of the *original* (classification) model. In theory, the softmax output represents the confidence of the model in assigning an input image to a particular class label, e.g. *Reject*, and may therefore be treated as a continuous quality score of some sort. However, the distribution of the quality scores as represented by the softmax output is qualitatively inferior, i.e. as expected, cross-entropy loss function biases softmax output towards the extremums, to those predicted by the regression model (Figure 1). Tuning the quality filter threshold using the softmax output would conceivably be less granular, e.g. setting the threshold to 0.4 or 0.5 would not make much of a difference to quality distribution, so we settled on the quality scores predicted by the regression model for all experiments.

Validation of the regression model. We first extracted the vascular network of 10,044 randomly selected Kaggle EyePACS retinal images using Deformable U-Net [8]. It is widely held that low image quality has a detrimental effect

on vessel segmentation [12,16]. As such, one would generally expect a smaller proportion of vascular network to be extracted from poorer images (Figure 2). In line with this, images predicted as having poorer quality tend to return a smaller proportion of vascular network (Pearson's r=-0.69; $p < 0.001$; see S4).

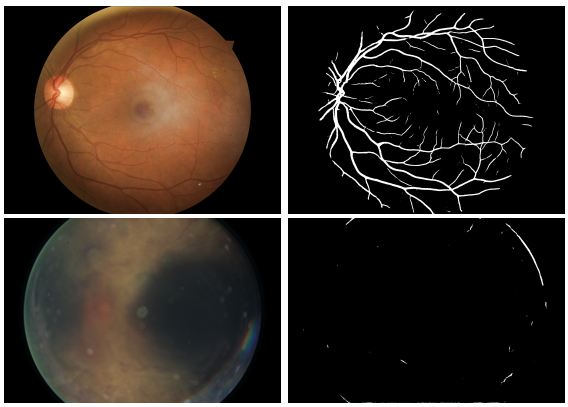

Fig. 2. Examples of images predicted as having superior quality (top left) and inferior quality (bottom left), with a quality score of 0.0 and 0.8, respectively, by the regression model. The extracted vascular network corresponding to each image is also displayed. A larger proportion of vascular network can be extracted from the high-quality image compared with the poor-quality one.

2.2 Effect of varying image quality threshold

DR is a common diabetic complication that affects the retina. Timely treatments are required to prevent or minimise vision loss when DR progresses to more severe stages, e.g. growth of new, leaky blood vessels. As such, many DL algorithms have been developed over the past few years to classify DR severity, with a view to aiding large-scale DR screening programmes [13]. The Kaggle EyePACS dataset was used in this study to elucidate the effect of varying quality threshold on the downstream DR classification task. Each image is labelled with an integer representing DR severity (ranges from 0 to 4) [21].

We should point out that the overwhelming majority of images graded as having the most advanced stage of DR (level 4) are of **poorer** quality (Figure 3; refer also to the figure in S2). This naturally leads one to wonder if excluding poorer training images might bias a downstream model against severe DR. The original dataset (n=88,702) was made up of a training set (40%) and a test set (60%). We used 30% (n=10,538) of training images to build a separate validation set. Images in each sub-dataset were filtered based on different pre-defined (either categorical or continuous) quality thresholds. A ResNet-50 pre-trained on ImageNet was then fine-tuned (detailed in S3) on the different resultant training sets, before comparing their test performance with one another.

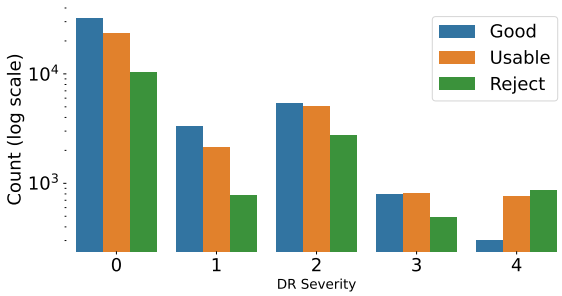

Fig. 3. Frequency of each quality label as predicted by the original MCF-Net across different DR severities. Similar figure using continuous score can be found in S2.

3 Experiments

3.1 Altering quality threshold on a categorical scale

As shown in Table 1, training the model on the *unfiltered (G+U+R)* training set consistently yielded the highest test accuracy across different combinations of quality labels – including the *unfiltered test set*. Furthermore, using *poorer images* on top of good images (*G+U+R*) gave rise to **optimal** performance on the test set comprised of *exclusively good images* (80.63%), which was even higher than the model trained exclusively on good images (79.08%). This is consistent with observation by Zhou et al. [23] that fine-tuning a model on poor images (originally trained on good images) did not hurt the model's performance on 'clean' data. Our observation therefore challenges conventional wisdom (see conclusion in [4]) that poor training images have undesirable effect on the test performance of a model, which presumably motivates the exclusion of suboptimal images in studies cited in Section 1.

Test \ Train	G+U+R (n=24,588)	G+U (n=20,458)	G+R (n=15,695)	U+R (n=13,023)	G (n=11,565)
G+U+R	78.24%	75.29%	76.11%	76.29%	75.72%
G+U	79.26%	76.74%	77.26%	77.34%	77.12%
G+R	78.66%	75.71%	76.63%	76.75%	76.36%
U+R	76.12%	72.49%	73.68%	74.11%	72.73%
G	**80.63%**	78.44%	78.85%	78.75%	79.08%
U	77.48%	74.53%	75.18%	75.48%	74.56%
R	73.25%	68.21%	70.53%	71.24%	68.89%

Table 1. Test accuracies (*row*) across different training sets (*column*). Best overall performance is highlighted in bold. *G='Good'; U='Usable'; R='Reject'.*

One caveat, however, is that the gain in performance accorded by the *G+U+R* training set in all test scenarios *might* arise largely as a result of its sheer size

(n=24,588) [17]. That said, the fact that a model trained on *U+R* – notwithstanding its small size (n=13,023) – still performs better on all test sets compared with a model trained on the much larger *G+U* training set (n=20,458), indicates that the observed difference in performance might still be attributable to a *variation in quality distribution* as opposed to the size of the training set.

Test \ Train	G+U+A (n=24,458)	G+R+A (n=24,195)	U+R+A (n=24,023)	G+A (n=24,130)
G+U+R	75.28%	74.51%	76.04%	73.80%
G+U	76.68%	75.77%	76.95%	74.93%
G+R	75.70%	75.19%	76.58%	74.65%
U+R	72.57%	71.72%	73.97%	70.95%
G	78.38%	77.66%	78.38%	76.98%
U	74.52%	73.30%	75.09%	72.24%
R	68.47%	68.41%	71.62%	68.23%

Table 2. Test accuracies (*row*) across different training sets (*column*), **augmented** such that the resultant number of training images matches that of G+U+R (n=24,588). *G='Good'; U='Usable'; R='Reject'; A='Augmentation'.*

What gives rise to the superior performance of *G+U+R*? Previous experiments were repeated after applying just *one of* the following conventional augmentation techniques to each randomly chosen training image: random rotation of no greater than 30 degrees, vertical flip, horizontal flip and Gaussian blur. 3 of these 4 techniques were not expected to alter image quality so the quality distribution between the original and augmented datasets could be assumed to be broadly similar. The number of augmented images was predefined such that the size of the resultant training set would be comparable to that of *G+U+R* (n=24,588), since our primary aim was to increase the size of the smaller training sets while preserving their quality distribution. Any difference in test performance between *G+U+R* could therefore be attributed largely to a variation in quality distribution.

Comparing the test performance of the model trained on *G+U+R* in Table 1 to that of the models trained on the different **augmented** training sets (Table 2), the former model still had a clear edge over all latter models. Importantly, *G+U+R*'s superior performance was evident across all test sets, *including those whose quality distribution differed from itself*. For instance, even after increasing the size of the *U+R* training set from 13,023 to 24,023, its performance on the *U+R* test set (73.97%) still lagged far behind that of the model trained on *G+U+R* (76.12%). The inferior performance of the augmented training sets vis-a-vis *G+U+R* training set in all test scenarios also lends credence to our proposition that conventional augmentations fail to capture the nuanced degradation in, and variation between, naturally acquired retinal images.

On a side note, the paradoxical observation that the *G+A* training set resulted in poorer test performance (e.g. 76.98% on *G* test set) than the *G* train-

ing set (79.08% on G test set) despite the former's significantly larger size could plausibly be the result of exacerbated class imbalance, which was already severe before augmentation (around 75% of G training images did not have DR). This unreservedly biased the model to level 0 DR (S5). However, it is unclear how augmentation could have worsened class imbalance as training images were randomly augmented with uniform probability. Taken together, our findings can only parsimoniously suggest that the optimally tuned training set does not owe its superior performance to its sheer size.

'Clean' training set biased model against severe DR. To see if training the model on exclusively good images would bias a model against more severe levels of DR (discussed in 2.2), we computed the model's accuracy for classifying images with level 3 and level 4 DR taken from the $G+U$ test set. As hypothesised, training the model on good images alone undermined its ability to classify level 3 DR (15.72%) and level 4 DR (0%) compared with the test performance gained by using the optimally tuned $G+U+R$ training set (36.12% and 24.00%, respectively). Augmenting the smaller training sets also did not improve their performance anywhere near that seen with $G+U+R$. This further justifies our contention against indiscriminate exclusion of poor training images and supports the notion of treating quality threshold as a tunable hyperparameter.

3.2 Altering quality threshold on a continuous scale

Quality threshold was gradually increased from 0.10 to 1.00, i.e. progressively poorer images are included at each step. All training details, e.g. model, optimiser, learning rate, seed for validation-test split, etc, were identical to previous experiments to allow for a fair comparison of results. The performance of the model trained on each resultant (filtered) training set was assessed based on its accuracy on two different test sets, i.e. *unfiltered* and *'Good'* (based on the original MCF-Net classification) images. As Figure 4 shows, the performance of the model tends to increase on both test sets with the inclusion of increasingly poorer images in the training set. The performance then peaked at 79.83% and 77.65% on the *'Good'* and *unfiltered* test sets, respectively, when the threshold was set to 0.72, and dropped from that point on.

This observation is consistent with our postulation about the presence of a *Goldilocks* level of image quality. Images beyond this optimal point are of such poor quality that they only serve to undermine the model. In support of this we observed *disproportionately* large changes in test accuracy as the threshold was changed from the optimal point to 1.00 and from 0.66 to the optimal point, i.e. -0.89% and +1.66%, despite relatively small changes in the number of images, i.e. +497 and +529 (see S6 for full table). Some *'Reject'* images were poorer still and had undesirable effect on test performance. Conversely, poor though those 529 additional images included at the optimal point were, they were beneficial insofar as they helped the model learn some 'usable' noise. As with before, the fact that adding increasingly poorer training images improved the model's

Fig. 4. Classification accuracy of the downstream model trained on different datasets filtered as per varying pre-defined continuous quality thresholds on two different test sets: *unfiltered* (blue) and *'good'* (categorical quality) images (orange).

performance on *good* images up until the optimal point contradicts conventional wisdom that using exclusively high-quality training images would yield optimal performance on high-quality test images. Taken together, the very presence of this optimal point which lies at some distance from the *expedient* threshold of 1.00 (i.e. inclusion of all images) *further strengthens* our justification that quality threshold can – and should indeed be – treated as a tunable hyperparameter.

3.3 Tuning on a continuous scale: does it confer *additional* value?

When quality threshold was tuned on a three-level categorical scale, the highest classification accuracy on the *unfiltered* and *'Good'* test sets came from the model trained on *G+U+R*, i.e. 78.24% and 80.63%, respectively (Table 1). If tuning the threshold on a *finer* scale had an *additional* benefit, one would expect the accuracy of the model on the same test sets to be even higher. However, the test accuracy from tuning the threshold on a *continuous* scale was in fact slightly lower – 77.40% and 79.83%, respectively (Figure 4). That said, our results **should not** be construed as evidence against the use of continuous over categorical scale because we had not been able to fully account for the stochastic nature of model training and evaluation, e.g. variation in minibatch images across runs, etc. This is evident if one considers the fact that the test accuracy of the model from the 'continuous' experiments with quality threshold set to 1.00 *did not agree with* its categorical equivalent (*G+U+R* training images).

4 Discussions and Conclusions

Considering the diminishing returns of increasing network complexity [22,2] and size of training data [18] in the domain of DL, it is apt that we focus our present work on the quality of input images. In particular, we propose – and have provided empirical justification – that image quality threshold should be treated as

a *tunable hyperparameter*. There is ample demonstration of the detrimental effect of synthetic image degradation on the performance of DL models trained on 'clean' datasets [1,9,15,7,20,4]. In line with this, natural sources of image degradation have also been shown to reduce the performance of a DL model trained exclusively on high-quality retinal images [22]. Our work is therefore driven by a desire to bring about a paradigm shift away from training a model exclusively on high-quality images to *carefully* curating a training set that also includes some suboptimal images. Indeed, when tested on poorer images (e.g. U test set) – in relation to the G test set – the G training set experienced the largest drop in accuracy among all training sets (Table 1).

To mitigate poor robustness to noise, much work has focused on retraining or fine-tuning an existing model with an augmented dataset – e.g. contrast reduction, Gaussian noise, defocus blur, etc. [9,20,23,5]. While these studies have unequivocally demonstrated an improvement in model's performance, this has only been demonstrated in *synthetically degraded* test images. It remains (largely) unclear how close such augmentations mimic naturally occurring degradation particular to retinal images, and if they can equally help a model generalise to such natural degradation as they are to synthetic degradation. Indeed, our concern is not unfounded because **even** generalisation across different types of *synthetic degradation* - from Gaussian noise to Gaussian blur [5] or from uniform defocus blur to oriented motion blur [20] - is not guaranteed. Our finding that the model trained on the *augmented* G training set did not have better performance on the poorer U test set than the model trained on the *original* G training set therefore *fills the aforementioned gap* by indicating that augmentation has *limited generalisability to naturally occurring degradation*. Our work also sets the scene for a solution centring on tuning the quality threshold for the training set.

Given that the stochastic nature of model training and evaluation has not been fully accounted for in this study, future studies could repeat each set of our experiments multiple times. This would allow one to better elucidate if there is any additional value in tuning the quality threshold on a *continuous* scale. Future work should also carry out a systematic investigation of the generalisability of other (more nuanced) augmentations such as contrast reduction, localised blur, etc. to naturally occurring degradation to help us confidently rule out the benefit of augmentation over inclusion of poor images. Moreover, other DL-based retinal image quality models could be used in addition to MCF-Net to verify the central thesis of this paper. To the best of our knowledge, we are the first to investigate the effect of tuning quality threshold on a downstream task related to retinal pathology. As we focused on DR, future work could make use of other retinal datasets [10], e.g. PALM, to see if similar conclusions apply to other diseases such as age-related macular degeneration.

References

1. Akkoca Gazioğlu, B.S., Kamaşak, M.E.: Effects of objects and image quality on melanoma classification using deep neural networks. Biomedical Signal Processing and Control **67**, 102530 (2021).

https://doi.org/https://doi.org/10.1016/j.bspc.2021.102530, bluehttps://www.sciencedirect.com/science/article/pii/S1746809421001270
2. Canziani, A., Paszke, A., Culurciello, E.: An analysis of deep neural network models for practical applications. CoRR **abs/1605.07678** (2016), bluehttp://arxiv.org/abs/1605.07678
3. da Costa, G.B.P., Contato, W.A., Nazare, T.S., Neto, J.E.S.B., Ponti, M.: An empirical study on the effects of different types of noise in image classification tasks (9 2016), bluehttp://arxiv.org/abs/1609.02781
4. Dodge, S., Karam, L.: Understanding how image quality affects deep neural networks (4 2016), bluehttp://arxiv.org/abs/1604.04004
5. Dodge, S.F., Karam, L.J.: Quality resilient deep neural networks. CoRR **abs/1703.08119** (2017), bluehttp://arxiv.org/abs/1703.08119
6. Fu, H., Wang, B., Shen, J., Cui, S., Xu, Y., Liu, J., Shao, L.: Evaluation of retinal image quality assessment networks in different color-spaces (7 2019). https://doi.org/10.1007/978-3-030-32239-7$_6$, bluehttp://arxiv.org/abs/1907.05345http://dx.doi.org/10.1007/978-3-030-32239-7_6
7. Jeelani, H., Martin, J., Vasquez, F., Salerno, M., Weller, D.S.: Image quality affects deep learning reconstruction of mri. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 357–360 (2018). https://doi.org/10.1109/ISBI.2018.8363592
8. Jin, Q., Meng, Z., Pham, T.D., Chen, Q., Wei, L., Su, R.: Dunet: A deformable network for retinal vessel segmentation. Knowledge-Based Systems **178**, 149–162 (Aug 2019). https://doi.org/10.1016/j.knosys.2019.04.025, bluehttp://dx.doi.org/10.1016/j.knosys.2019.04.025
9. Kugelman, J., Alonso-Caneiro, D., Read, S.A., Vincent, S.J., Chen, F.K., Collins, M.J.: Effect of altered oct image quality on deep learning boundary segmentation. IEEE Access **8**, 43537–43553 (2020). https://doi.org/10.1109/ACCESS.2020.2977355
10. Li, T., Bo, W., Hu, C., Kang, H., Liu, H., Wang, K., Fu, H.: Applications of deep learning in fundus images: A review. Medical Image Analysis **69**, 101971 (2021). https://doi.org/https://doi.org/10.1016/j.media.2021.101971, bluehttps://www.sciencedirect.com/science/article/pii/S1361841521000177
11. Lin, G.M., Chen, M.J., Yeh, C.H., Lin, Y.Y., Kuo, H.Y., Lin, M.H., Chen, M.C., Lin, S.D., Gao, Y., Ran, A., Cheung, C.Y.: Transforming retinal photographs to entropy images in deep learning to improve automated detection for diabetic retinopathy. Journal of Ophthalmology **2018** (2018). https://doi.org/10.1155/2018/2159702
12. Moccia, S., De Momi, E., El Hadji, S., Mattos, L.S.: Blood vessel segmentation algorithms — review of methods, datasets and evaluation metrics. Computer Methods and Programs in Biomedicine **158**, 71–91 (2018). https://doi.org/https://doi.org/10.1016/j.cmpb.2018.02.001, bluehttps://www.sciencedirect.com/science/article/pii/S0169260717313421
13. Ng, W., Zhang, S., Wang, Z., Ong, C., Gunasekeran, D., Lim, G., Zheng, F., Tan, S., Tan, G., Rim, T., Schmetterer, L., Ting, D.: Updates in deep learning research in ophthalmology. Clinical Science **135**(20), 2357–2376 (10 2021). https://doi.org/10.1042/CS20210207, bluehttps://doi.org/10.1042/CS20210207
14. Poplin, R., Varadarajan, A.V., Blumer, K., Liu, Y., McConnell, M.V., Corrado, G.S., Peng, L., Webster, D.R.: Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nature Biomedical Engineering **2**, 158–164 (3 2018). https://doi.org/10.1038/s41551-018-0195-0

15. RichardWebster, B., Anthony, S.E., Scheirer, W.J.: Psyphy: A psychophysics driven evaluation framework for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(9), 2280–2286 (2019). https://doi.org/10.1109/TPAMI.2018.2849989
16. Singh, N., Kaur, L.: A survey on blood vessel segmentation methods in retinal images. In: 2015 International Conference on Electronic Design, Computer Networks Automated Verification (EDCAV). pp. 23–28 (2015). https://doi.org/10.1109/EDCAV.2015.7060532
17. Sun, C., Shrivastava, A., Singh, S., Gupta, A.: Revisiting unreasonable effectiveness of data in deep learning era. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
18. Sun, C., Shrivastava, A., Singh, S., Gupta, A.: Revisiting unreasonable effectiveness of data in deep learning era. CoRR **abs/1707.02968** (2017), bluehttp://arxiv.org/abs/1707.02968
19. Varadarajan, A.V., Poplin, R., Blumer, K., Angermueller, C., Ledsam, J., Chopra, R., Keane, P.A., Corrado, G.S., Peng, L., Webster, D.R.: Deep learning for predicting refractive error from retinal fundus images. Investigative Ophthalmology and Visual Science **59**, 2861–2868 (6 2018). https://doi.org/10.1167/iovs.18-23887
20. Vasiljevic, I., Chakrabarti, A., Shakhnarovich, G.: Examining the impact of blur on recognition by convolutional networks. CoRR **abs/1611.05760** (2016), bluehttp://arxiv.org/abs/1611.05760
21. Wilkinson, C.P., Ferris, F.L., Klein, R.E., Lee, P.P., Agardh, C.D., Davis, M., Dills, D., Kampik, A., Pararajasegaram, R., Verdaguer, J.T., Lum, F.: Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology **110**, 1677–1682 (9 2003). https://doi.org/10.1016/S0161-6420(03)00475-5
22. Yip, M., Lim, G., Lim, Z., Nguyen, D.Q., Chong, C., Yu, M., Bellemo, V., Xie, Y., Lee, X., Hamzah, H., Ho, J., Tan, T.E., Sabanayagam, C., Grzybowski, A., Tan, G., Hsu, W., Lee, M., Wong, T.Y., Ting, D.: Technical and imaging factors influencing performance of deep learning systems for diabetic retinopathy screening. npj Digital Medicine **3** (03 2020). https://doi.org/10.1038/s41746-020-0247-1
23. Zhou, Y., Song, S., Cheung, N.: On classification of distorted images with deep convolutional neural networks. CoRR **abs/1701.01924** (2017), bluehttp://arxiv.org/abs/1701.01924