Lightweight Deep Learning Models for Diabetic Retinopathy Grading: A Cross Dataset Generalization and Edge Deployment Study

Muhammad Adeel Abid, Naina Noor Sindhu, Faiqa Tanveer
Department of Data Science & AI (Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan).
*Correspondence: adeel.abid@kfueit.edu.pk

Citation | Abid. M. A, Sindhu. N. N, Tanveer. F, "Lightweight Deep Learning Models for Diabetic Retinopathy Grading: A Cross Dataset Generalization and Edge Deployment Study", IJIST, Vol. 8 Issue. 10 pp 3992-4011, November 2026
Received | Oct 12, 2026 **Revised |** Oct 27, 2026 **Accepted |** Oct 30, 2026 **Published |** Nov 01, 2026.

Diabetic retinopathy (DR) is a leading cause of preventable blindness. Automated grading of fundus photographs could expand screening where eye specialists are few. Few studies test lightweight models on a second dataset or under deployment limits. This study compares five lightweight networks (MobileNetV3, EfficientNet Lite0, ShuffleNetV2, GhostNet and MobileViT XS) and a ResNet50 reference on APTOS 2019 and EyePACS. Duplicates are removed and each model is trained with three seeds. Same dataset quadratic weighted kappa (QWK) is 0.865 to 0.895 on APTOS and 0.420 to 0.533 on EyePACS. Cross dataset QWK falls to 0.162 to 0.342 from APTOS to EyePACS and 0.350 to 0.653 in the other direction. For referable DR the models detect only 14% to 23% of referable cases from APTOS to EyePACS (specificity 0.95 to 0.98). From EyePACS to APTOS sensitivity is 0.87 to 0.99 but specificity is 0.36 to 0.72. MobileViT XS gives the highest mean QWK in three of four settings. Adding about 10 percent labelled target images raises its cross dataset QWK to 0.506 and 0.823. Controls show that this gain comes from target images and not from more training data. INT8 quantization made files 2.8 to 3.9 times smaller. It keeps QWK for ShuffleNetV2, EfficientNet Lite0 and ResNet50 (loss at most 0.048) but not for MobileNetV3, GhostNet and MobileViT XS (loss 0.217 to 0.478). Percentile calibration reduces this loss but does not remove it. On a 4-core CPU the fp32 median latency is 6.9 to 113.8 ms per image and peak memory is 86 to 293 MB. Latency is not measured on edge hardware. Lightweight models need cross dataset and deployment tests before screening use.

Keywords: Diabetic Retinopathy; Lightweight Deep Learning; Domain Shift; INT8 Quantization; Edge Deployment; Grad CAM.

Introduction:

Diabetes is common and its burden is growing. One estimate gives 536.6 million adults aged 20 to 79 with diabetes in 2021 and 783.2 million by 2045 [1]. Among people with diabetes the global prevalence of DR is 22.27 percent and that of vision threatening DR is 6.17 percent [2]. The same review projects 160.5 million adults with DR by 2045 [2]. Color fundus photographs allow screening for DR. Deep learning has been widely applied to detect and grade the disease. A recent review of 91 studies and 23 public datasets found that deep learning generally outperforms traditional methods [3].

A model that works on one image collection can fail on another. Machine learning for medical images suffers from domain shift when training and target data follow different distributions [4]. In DR, the shift can arise from differences between clinical centres, equipment and patients [5]. Che et al. describe three generalization issues in DR grading: visual-style and image-degradation shifts, diagnostic pattern diversity and data imbalance [6]. Atwany and Yaqub report that models trained on a single source can underperform on other DR datasets [7]. Several methods address the problem. They include DRGen [7], mutual information regularization [5], variational autoencoders [8] and the GDRNet framework [6]. A survey covers domain generalization for medical images more broadly [9].

Screening in rural and crowded regions may rely on mobile systems [10]. Esengönül et al. tested five lightweight architectures for mobile DR detection. They found that dataset inclusivity and size strongly affect accuracy and sensitivity, and that EfficientNet based models did best on APTOS [10]. [11] deployed MobileNet, ShuffleNet, SqueezeNet and a custom network on smartphones and microcontrollers and reported inference speed and resource use. Lightweight models matter for DR screening because clinics and mobile units often have limited computing, power and network access. Small models need less memory and can run on the device. They also cost less to deploy at scale. These gains bring challenges. A small model has less capacity for fine lesion detail. Its accuracy may fall further on a new camera or population [5][6]. Compression such as INT8 quantization can add more loss, and the loss differs by architecture [12][13].

Quantization lowers numerical precision to cut memory and computation. It matters when resources are restricted [12]. Methods built for convolutional networks do not suit attention layers well, so vision transformers need other treatment such as ranking awareness and mixed precision schemes [13].

Grad CAM [14] shows image regions that drive a prediction. Ghassemi et al. argue that current explainability methods are unlikely to deliver trust for patient level decisions. They call for rigorous internal and external validation [15]. Dey et al. note that DR grading is limited by the availability of high-quality datasets [16].

Most of these studies propose a new method or test one family of models. Fewer put several common lightweight networks through one protocol. Fewer still ask what a shift between two datasets does to each network, whether simple fixes help, and what happens when the same networks are quantized. This study does these things together. It uses duplicate removal, and patient level splitting for EyePACS. It reports three seeds with confidence intervals. It ends with a deployment test and a look at Grad CAM maps of failures.

The literature leaves four gaps. First, DR domain generalization methods usually use one large backbone [5][6][7]. Few studies test simple fixes across several lightweight networks. Second, cross dataset results are often given in one direction and without sensitivity and specificity for referable DR. Third, work on quantizing vision transformers [13] is not linked to DR grading, and mobile DR studies [10][11] do not compare how INT8 changes accuracy across architectures. Fourth, clinical use needs external validation and tested explanations [15]. This study addresses the first three gaps on two public datasets. It does not address the fourth.

Objective of the Study:

The specific objectives are as follows:

Compare five lightweight networks and ResNet50 on APTOS and EyePACS.

Measure the loss when each network is tested on the other dataset.

Test color jitter, color normalization and target image mixing, with equal size controls.

Measure the effect of INT8 on size, latency, memory and QWK, and test calibration options.

Inspect Grad CAM maps of cross dataset failures.

Objectives 1 and 2 are covered in Models, Evaluation and the same dataset and cross dataset results. Objective 3 is covered in Mitigation methods and the mitigation results. Objective 4 is covered in Edge deployment and its results. Objective 5 is covered in Grad CAM and its results.

Novelty Statement:

Earlier work on DR domain shift proposes training methods and usually uses larger networks [5][6][7]. Earlier work on mobile DR models reports accuracy or speed on one dataset or device [10][11]. As far as we know, no study runs several common lightweight networks through one protocol that combines duplicate removal, patient level splitting for EyePACS, two-way cross dataset tests, mitigation controls and INT8 quantization. This study makes three contributions. First, a ResNet50 reference does not generalize better than small networks. Second, an equal size control shows that the gain from adding target images comes from target exposure and not extra data. Third, the network that generalizes best (MobileViT XS) loses most of its QWK under INT8, and percentile calibration only partly restores it.

Materials and Methods:

Investigation site:

This study has no physical field site. The data comes from two public Kaggle collections of color fundus photographs. They are APTOS 2019 Blindness Detection [17] and the EyePACS Diabetic Retinopathy Detection set [18]. Both use the same five grades of DR: 0 no DR, 1 mild, 2 moderate, 3 severe and 4 proliferative. The full competition files are large. Smaller resized copies shared by Kaggle users are used instead. We chose APTOS 2019 and EyePACS because both are public, use the same five grade scale and come from different sources. This suits a test of dataset shift. We did not remove images for poor quality. Labels: We used the labels as provided and did not regrade them. In 30 of 123 duplicate groups in APTOS (24 percent) the copies carried different grades. This shows label noise. Duplicates. We found exact duplicates with MD5 hashes of the processed 224 by 224 arrays. We did not search for near duplicates. The APTOS copy is the 224 by 224 color version of the 2019 training images [19]. The EyePACS copy is the plain resized training set [20]. The cropped version of that set is not used. The resizing is done by the uploaders and is not independently verified. Figure 1 gives an overview of the study.

The APTOS training set holds 3,662 labelled images. Exact duplicates are found by comparing MD5 hashes of the processed 224 by 224 arrays. This gave 123 groups of repeated images. In 30 of these groups the copies carried different grades. All copies in those 30 groups are removed. In the other groups one copy is kept. This left 3,504 images. Without this step identical images appeared in more than one split. APTOS has no patient identifiers, so a patient level split is not possible. Images of one patient may appear in more than one split. This may raise APTOS same dataset QWK. We reduced the risk by removing exact duplicates. APTOS to EyePACS results is not affected because the EyePACS test split is patient level. EyePACS to APTOS results with no mixing are also not affected, because those models never saw APTOS. The EyePACS training set holds 35,126 images from both eyes of patients. The patient number is part of each file name. A random sample of about 28.5 percent of patients is taken. It is stratified by the worst eye grade. This gave 10,010 images. The patients are then split 70/15/15 into training, validation and test sets. Both eyes of a patient always stay in the

same split. No patient and no identical image appear in two splits. Table 1 gives the class counts.

Table 1. Number of images per grade in each split.

Dataset	Split	Images	No DR	Mild	Moderate	Severe	Proliferative
APTOS	Train	2,452	1,257	236	645	124	190
APTOS	Val	526	269	51	139	26	41
APTOS	Test	526	270	51	138	27	40
EyePACS	Train	7,006	5,174	471	1,054	166	141
EyePACS	Val	1,502	1,107	98	228	38	31
EyePACS	Test	1,502	1,103	101	230	38	30

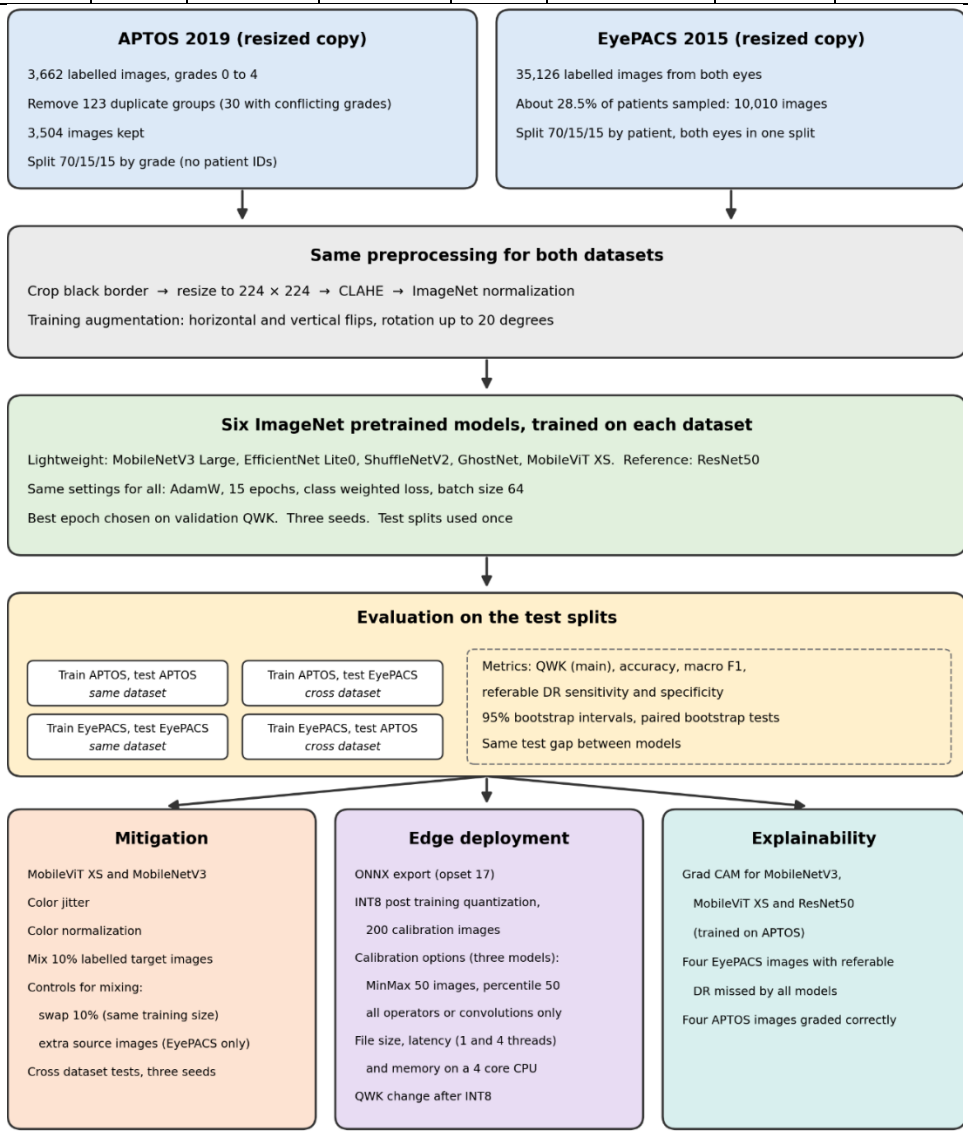


Figure 1. Overview of the study design.

Figure 1 shows seven steps. (1) Dataset preparation: two Kaggle copies are loaded, exact duplicates are removed, APTOS is split by grade and EyePACS by patient. (2) Preprocessing: border crop, resize to 224 by 224, CLAHE and ImageNet normalization. (3) Training: six pretrained models, three seeds, 15 epochs, selection by validation QWK. (4) Evaluation: same dataset and cross dataset QWK, sensitivity, specificity and bootstrap intervals. (5) Mitigation: color jitter, color normalization and target image mixing with two size

controls. (6) Quantization: ONNX export, INT8 with calibration options, CPU latency and memory. (7) Explainability: Grad CAM on APTOS trained models.

Preprocessing and Augmentation:

The same pipeline is applied to both datasets. Black borders are cropped by taking the bounding box of pixels with a grey value above 10. Images are resized to 224 by 224 pixels with area interpolation. Contrast limited adaptive histogram equalization (CLAHE) [21] is applied to the lightness channel in LAB color space with a clip limit of 2.0 and 8 by 8 tiles. Images are then normalized with the ImageNet means and standard deviations. Training used random horizontal flips, random vertical flips and random rotation up to 20 degrees.

Models:

Five lightweight models and one larger reference model are compared. The lightweight models are MobileNetV3 Large [22], EfficientNet Lite0 (a lighter member of the EfficientNet family) [23], ShuffleNetV2 1.0x [24], GhostNet 1.0x [25] and MobileViT XS [26]. ResNet50 [27] is the reference. All models started from ImageNet pretrained weights and received a new output layer with five classes. FLOPs are counted for one 224 by 224 image with the PyTorch FLOP counter. One multiply add counts as two FLOPs. Table 2 gives the size and compute of each model.

Table 2. Size and compute of the six models at 224 by 224 input.

Model	Parameters (millions)	GFLOPs	File size fp32 (MB)
MobileNetV3	4.21	0.43	17.0
EfficientNet Lite0	3.38	0.77	13.8
ShuffleNetV2	1.26	0.29	5.2
GhostNet	3.91	0.28	15.9
MobileViT XS	1.93	1.41	7.9
ResNet50	23.52	8.17	94.4

Training:

Every model used the same settings. The optimizer is AdamW with a learning rate of 0.0003 and weight decay of 0.0001. The learning rate followed a cosine schedule over 15 epochs. The batch size is 64. The loss is cross entropy with class weights set to the inverse of the class frequency in the training split. Training used mixed precision. The epoch with the highest QWK on the validation split of the training dataset is kept. Each model is trained with three seeds (0, 1 and 2) on each dataset. Training ran on a Kaggle notebook with one NVIDIA T4 GPU. Software/libraries used are Python, PyTorch, torchvision, timm and ONNX Runtime 1.31.0. Pretrained weights came from timm (mobilenetv3_large_100, mobilevit_xs, ghostnet_100) and torchvision (ShuffleNetV2, IMAGENET1K_V1). The batch loader used 2 workers and dropped the last incomplete batch. Seeds are set for PyTorch and NumPy. The checkpoint with the highest validation QWK is kept.

Evaluation:

Same dataset results come from training and testing on the same dataset. Cross dataset results come from training on one dataset and testing on the test split of the other. The test splits are used once at the end. Model selection used the validation split only, which follows the call for a correct model selection strategy in DR domain generalization work [5] and in general domain generalization work [28]. For one model the drop is its dataset QWK minus its cross dataset QWK. All models score higher on the APTOS test split than on the EyePACS test split, so this drop can be negative. A second measure is therefore used, called the same test gap. For one test split it is the QWK of the model trained on that dataset minus the QWK of the model trained on the other dataset. The gap is computed per seed and averaged. The mean same test gap is the average over both test splits. The main metric is quadratic weighted kappa (QWK). Accuracy and macro F1 are also reported. Sensitivity and specificity are

computed for referable DR, defined as grade 2 or higher. For each test split we drew 1,000 bootstrap resamples of the test images with replacement. In each resample we computed QWK for each of the three seeds and averaged them. The 2.5th and 97.5th percentiles gave the 95 percent interval. To compare two models, we used the same resamples for both and took the difference in each resample. A difference is called significant when the 95 percent interval of the difference excludes zero. We made 20 comparisons and did not adjust for them. We treat these comparisons as exploratory. Only differences with wide margins should be taken as firm.

Mitigation Methods:

Three methods for reducing the cross-dataset performance drop were tested on MobileViT XS and MobileNetV3 with the same training settings and seeds. Color jitter added random changes of brightness, contrast and saturation of up to 0.4 and hue of up to 0.05 to the augmentation. Color normalization standardized each color channel inside the retina to a mean of 110 and a standard deviation of 40 before CLAHE. Mixing added about 10 percent of the labelled training split of the other dataset to the training data. This is about 700 EyePACS images or 245 APTOS images. They are drawn at random by grade with a different draw for each seed. Model selection still used the validation split of the main training dataset only. The baseline for each comparison is the result in Table 3. Two controls test whether the gain comes from extra images or from target images. In the swap control, 10 percent of the training split is replaced by the same number of target images, so training size equals the baseline. In the extra source control, the same number of extra source images from unused EyePACS patients is added. The unused pool has 25,116 images from 12,558 patients and no patient overlaps with any split. This control is not possible for APTOS because every image is used. Both controls use the same target draw as mixing, with MobileViT XS and MobileNetV3 and three seeds.

Edge Deployment:

The weights of the best epoch of seed 0 are exported to ONNX (opset 17). Post training static INT8 quantization used ONNX Runtime in QDQ format with uint8 activations and per channel int8 weights. Calibration used 200 images from the training split and the default ONNX Runtime calibration method. To test calibration options, MobileViT XS, MobileNetV3 and GhostNet (seed 0) are also quantized with MinMax on 50 images and with percentile calibration on 50 images [29]. Each option is applied to all operators or to convolutions only. The percentile method used 50 images because of memory limits. Calibration images are drawn at random from the training split. QWK is measured for the fp32 and INT8 ONNX models on both test splits. The fp32 ONNX models reproduced PyTorch in dataset QWK within 0.003. Latency is measured with the ONNX Runtime CPU provider on a Kaggle notebook CPU with 4 cores. One 224 by 224 image is processed at a time. Ten warm up runs are followed by 100 timed runs and the median reported. Latency is measured with 1 thread and with 4 threads. Memory is the peak resident memory of the benchmark process. The memory of the runtime itself (about 47.5 MB) is measured after importing ONNX Runtime and subtracted to give the model memory. Each measurement ran in its own process. Latency and memory in Table 7 are the means of the two trained variants of each model. We had no access to an edge device or a power meter. The CPU results are a proxy for edge performance. Power use is not measured.

Grad CAM:

Grad CAM [14] is computed for MobileNetV3, MobileViT XS and ResNet50 trained on APTOS (seed 0). The target layer is the last feature block of MobileNetV3, the final convolution of MobileViT XS and the last stage of ResNet50. The target class is the predicted class. Four EyePACS test images with referable DR are chosen at random from those that all

three models graded below 2. Four APTOS test images with referable DR that all three models graded exactly right are chosen in the same way. The random seed is fixed.

Results and Discussion:

Color Differences Between the Datasets:

Retinal pixels differed in color between the datasets. In a sample of up to 1,000 training images per dataset taken before CLAHE the mean red, green and blue values are 136.9, 73.8 and 24.0 for APTOS and 131.0, 91.4 and 65.8 for EyePACS. The green and blue channels differ most.

In Dataset and Cross Dataset Performance:

On APTOS all six models reached an in dataset QWK between 0.865 and 0.895 (Table 3). ResNet50 had the highest mean (0.895). Its lead over MobileViT XS (difference 0.001) and over GhostNet (0.011) is not significant. It is significantly higher than MobileNetV3, EfficientNet Lite0 and ShuffleNetV2 by 0.024 to 0.030. On EyePACS the in dataset QWK is lower at 0.420 to 0.533. MobileViT XS is the best model and is significantly above all others by 0.045 to 0.112.

Cross dataset results are much lower (Table 3 and Figure 2). Trained on APTOS and tested on EyePACS the models reached 0.162 to 0.342. MobileViT XS is highest. Its lead over MobileNetV3 (0.299) is not significant (difference 0.041, interval from −0.003 to 0.086). It is significantly above the other four models by 0.067 to 0.178. ResNet50 is lowest at 0.162. Trained on EyePACS and tested on APTOS the models reached 0.350 to 0.653. MobileViT XS is highest and significantly above all others by 0.047 to 0.302. ShuffleNetV2 is lowest at 0.350. The full intervals and the pairwise tests are in Tables 10 and 11 (Appendix).

When training on APTOS the drop from in dataset to cross dataset QWK is 0.552 for MobileViT XS, 0.566 for MobileNetV3, 0.597 for ShuffleNetV2, 0.655 for GhostNet, 0.664 for EfficientNet Lite0 and 0.732 for ResNet50. The mean same test gap is lowest for MobileNetV3 (0.213) and MobileViT XS (0.216) and highest for ResNet50 (0.333) and ShuffleNetV2 (0.348). Seed to seed variation is small in dataset (SD 0.004 to 0.030). It is large for EyePACS to APTOS for ResNet50 (0.127), GhostNet (0.116) and ShuffleNetV2 (0.095). Accuracy hides the cross-dataset problem. From APTOS to EyePACS accuracy is 0.675 to 0.722 but macro F1 is only 0.239 to 0.278. About 73 percent of the EyePACS test images are grade 0.

Table 3. QWK (mean ± SD over three seeds) for each model and the mean same test gap. Drops and gaps are computed from unrounded means. SD uses n minus 1.

Model	Train APTOS, test APTOS	Train APTOS, test EyePACS	Train EyePACS, test EyePACS	Train EyePACS, test APTOS	Mean same test gap
MobileNetV3	0.865 ± 0.004	0.299 ± 0.026	0.465 ± 0.024	0.606 ± 0.052	0.213
EfficientNet Lite0	0.867 ± 0.005	0.203 ± 0.033	0.420 ± 0.004	0.562 ± 0.008	0.261
ShuffleNetV2	0.871 ± 0.016	0.275 ± 0.032	0.449 ± 0.030	0.350 ± 0.095	0.348
GhostNet	0.883 ± 0.010	0.229 ± 0.003	0.443 ± 0.009	0.583 ± 0.116	0.257
MobileViT XS	0.894 ± 0.005	0.342 ± 0.028	0.533 ± 0.015	0.653 ± 0.044	0.216
ResNet50	0.895 ± 0.007	0.162 ± 0.017	0.489 ± 0.011	0.556 ± 0.127	0.333

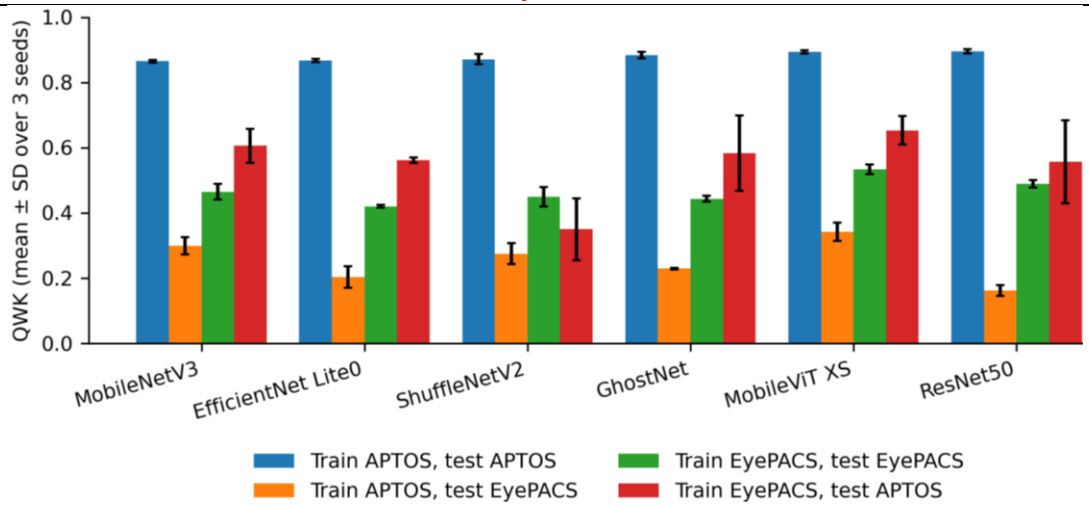


Figure 2. QWK for the six models in the four train and test settings. Bars show the mean over three seeds and error bars show one standard deviation.

Referable DR:

Trained on APTOS and tested on EyePACS the models found only 14 to 23 percent of referable cases. Specificity stayed at 0.95 to 0.98 (Table 4). Trained on EyePACS and tested on APTOS sensitivity is 0.87 to 0.99 but specificity is only 0.36 to 0.72. MobileViT XS had the best balance in that direction (sensitivity 0.989 and specificity 0.718). In dataset values are in Table 10 (Appendix).

Table 4. Sensitivity and specificity for referable DR (grade 2 or higher) in the two cross dataset settings. Values are means over three seeds.

Model	APTOS to EyePACS sensitivity	APTOS to EyePACS specificity	EyePACS to APTOS sensitivity	EyePACS to APTOS specificity
MobileNetV3	0.227	0.958	0.896	0.646
EfficientNet Lite0	0.164	0.957	0.865	0.631
ShuffleNetV2	0.163	0.977	0.967	0.356
GhostNet	0.141	0.980	0.956	0.555
MobileViT XS	0.220	0.962	0.989	0.718
ResNet50	0.147	0.954	0.977	0.512

Mitigation:

Mixing about 10 percent of labelled target images is the only method that improved all four cross dataset tests in all three seeds (Table 5 and Figure 3). For MobileViT XS QWK rose from 0.342 to 0.506 from APTOS to EyePACS and from 0.653 to 0.823 from EyePACS to APTOS. For MobileNetV3 it rose from 0.299 to 0.380 and from 0.606 to 0.814. For MobileViT XS from APTOS to EyePACS sensitivity rose from 0.220 to 0.547 and specificity fell from 0.962 to 0.907. Computed from the seed means the same test gap of MobileViT XS fell from 0.241 to 0.078 on APTOS and from 0.191 to 0.064 on EyePACS. For MobileNetV3 it fell from 0.259 to 0.070 and from 0.166 to 0.116. Mixing also raised the in dataset QWK in all four cases by 0.007 to 0.038 (Table 12, Appendix). Two controls test this (Table 6). Replacing source images by the same number of target images gave almost the same cross dataset QWK as mixing: 0.506 against 0.506 and 0.811 against 0.823 for MobileViT XS, and 0.350 against 0.380 and 0.805 against 0.814 for MobileNetV3. Adding extra source images did not help (−0.008 for MobileNetV3 and +0.029 for MobileViT XS from EyePACS to APTOS). The same dataset gain seen with mixing did not appear when the training size is held equal (−0.015 to +0.013). The cross dataset gain therefore comes from the target images and not from extra training data. Even with mixing sensitivity from APTOS to EyePACS reached only 0.513 for MobileNetV3 and 0.547 for MobileViT XS.

Color jitter gave small gains from APTOS to EyePACS (+0.045 for MobileNetV3 and +0.023 for MobileViT XS). From EyePACS to APTOS the results are mixed (−0.038 for MobileNetV3 and +0.073 for MobileViT XS with a large seed spread). Color normalization did not help. It lowered the EyePACS to APTOS QWK of both models in all three seeds (−0.161 and −0.044). In the other direction it changed QWK by +0.007 and +0.014.

Table 5. Cross dataset QWK, change against baseline, number of seeds with an improvement, and referable DR sensitivity and specificity for each mitigation method (mean ± SD over three seeds). Changes are computed from unrounded means. SD uses n minus 1.

Model	Direction	Method	QWK	Change	Seeds better	Sens.	Spec.
MobileViT XS	APTOS to EyePACS	Baseline	0.342 ± 0.028	-	-	0.220	0.962
MobileViT XS	APTOS to EyePACS	Color jitter	0.364 ± 0.011	+0.023	2 of 3	0.304	0.938
MobileViT XS	APTOS to EyePACS	Color normalization	0.356 ± 0.017	+0.014	2 of 3	0.326	0.934
MobileViT XS	APTOS to EyePACS	Mix 10% target	0.506 ± 0.011	+0.165	3 of 3	0.547	0.907
MobileViT XS	EyePACS to APTOS	Baseline	0.653 ± 0.044	-	-	0.989	0.718
MobileViT XS	EyePACS to APTOS	Color jitter	0.725 ± 0.083	+0.073	2 of 3	0.971	0.810
MobileViT XS	EyePACS to APTOS	Color normalization	0.609 ± 0.020	-0.044	0 of 3	0.959	0.728
MobileViT XS	EyePACS to APTOS	Mix 10% target	0.823 ± 0.007	+0.170	3 of 3	0.907	0.901
MobileNetV3	APTOS to EyePACS	Baseline	0.299 ± 0.026	-	-	0.227	0.958
MobileNetV3	APTOS to EyePACS	Color jitter	0.344 ± 0.010	+0.045	3 of 3	0.311	0.925
MobileNetV3	APTOS to EyePACS	Color normalization	0.306 ± 0.045	+0.007	2 of 3	0.310	0.917
MobileNetV3	APTOS to EyePACS	Mix 10% target	0.380 ± 0.018	+0.082	3 of 3	0.513	0.842
MobileNetV3	EyePACS to APTOS	Baseline	0.606 ± 0.052	-	-	0.896	0.646
MobileNetV3	EyePACS to APTOS	Color jitter	0.568 ± 0.034	-0.038	1 of 3	0.956	0.574
MobileNetV3	EyePACS to APTOS	Color normalization	0.444 ± 0.058	-0.161	0 of 3	0.676	0.747
MobileNetV3	EyePACS to APTOS	Mix 10% target	0.814 ± 0.019	+0.208	3 of 3	0.893	0.886

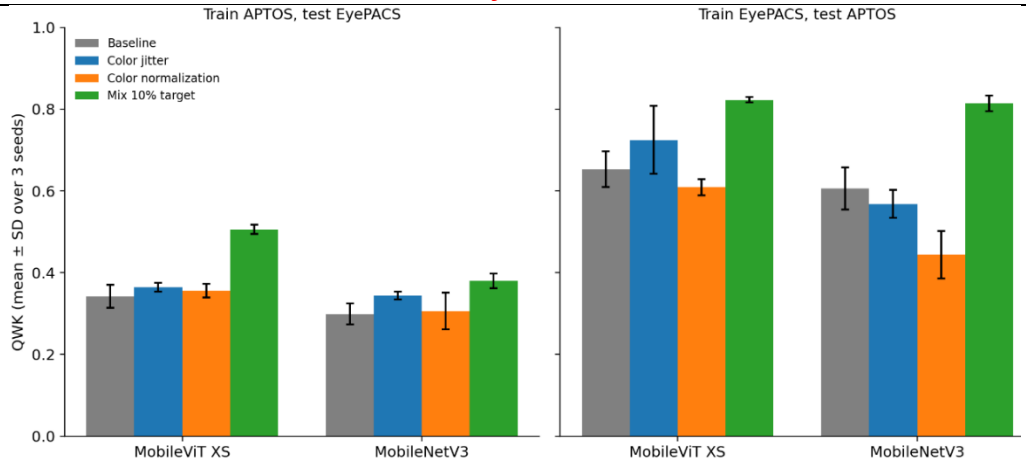


Figure 3. Cross dataset QWK of the baseline and the three mitigation methods. Bars show the mean over three seeds and error bars show one standard deviation.

Table 6. Cross dataset QWK (mean over three seeds) for mixing and the two size controls.

Model, direction	Baseline	Mix 10%	Swap 10%	Extra source
MobileViT XS, APTOS to EyePACS	0.342	0.506	0.506	not possible
MobileViT XS, EyePACS to APTOS	0.653	0.823	0.811	0.682
MobileNetV3, APTOS to EyePACS	0.299	0.380	0.350	not possible
MobileNetV3, EyePACS to APTOS	0.606	0.814	0.805	0.598

Edge Deployment:

INT8 quantization made the files 2.8 to 3.9 times smaller (Table 7). The fp32 ONNX models are 5.1 to 94.0 MB and the INT8 models are 1.6 to 24.0 MB. ResNet50 had the largest memory saving (peak memory 293 MB to 121 MB). INT8 kept QWK close to fp32 for ShuffleNetV2 (in dataset change of 0.007 and 0.032 for the APTOS and EyePACS trained models), EfficientNet Lite0 (0.020 and 0.048) and ResNet50 (−0.004 and 0.046). It lowered QWK strongly for MobileNetV3 (0.478 and 0.270), GhostNet (0.237 and 0.217) and MobileViT XS (0.398 and 0.427). Table 8 and Figure 4 show the details. For MobileViT XS trained on EyePACS the cross dataset QWK on APTOS fell from 0.607 to 0.063. Table 9 compares calibration options. Percentile calibration gave the smallest loss in five of six cases, but no option restored the fp32 same dataset QWK in any of the six cases. Calibration did not change latency (about 12 ms for MobileNetV3 and 41 ms for MobileViT XS). Convolution only versions were slower, and larger for MobileViT XS. Repeating the default setting with a different calibration draw changed same dataset QWK by up to 0.07 and cross dataset QWK by up to 0.09. Differences below about 0.07 should not be trusted.

On this CPU INT8 is faster than fp32 only for EfficientNet Lite0 (16.1 ms to 12.4 ms with one thread) and ResNet50 (113.8 ms to 69.3 ms). For MobileNetV3, GhostNet and MobileViT XS it is slower. For ShuffleNetV2 it is equal. MobileViT XS with INT8 did not become faster with more threads (42.9 ms with one thread and 42.3 ms with four). The fastest model is ShuffleNetV2 at 6.9 ms in fp32. MobileNetV3 ran in 10.0 ms in fp32 and MobileViT XS in 38.5 ms. Latency does not follow compute alone. GhostNet has the lowest compute (0.28 GFLOPs) but is slower than ShuffleNetV2 (0.29 GFLOPs), at 11.4 ms against 6.9 ms. Parameters and GFLOPs (Table 2) do not depend on hardware but predict latency only loosely.

Table 7. File size, median latency per image and memory of the fp32 and INT8 ONNX models on a 4 core Kaggle CPU. Values are means of the two trained variants of each model.

Model	Size (MB)		Latency 1 thread (ms)		Latency 4 threads (ms)		Peak memory (MB)		Model memory (MB)	
Model	**fp32**	**INT 8**	**fp32**	**INT 8**	**fp32**	**INT 8**	**fp32**	**INT 8**	**fp32**	**INT 8**
MobileNetV3	16.8	4.7	10.0	12.5	5.7	9.4	112	81	65	33
EfficientNet Lite0	13.5	3.8	16.1	12.4	8.1	6.8	105	78	58	30
ShuffleNetV2	5.1	1.6	6.9	6.8	5.1	5.8	86	77	38	29
GhostNet	15.7	4.5	11.4	12.3	8.5	9.3	110	83	63	35
MobileViT XS	7.9	2.8	38.5	42.9	24.9	42.3	107	90	59	42
ResNet50	94.0	24.0	113.8	69.3	55.5	35.6	293	121	245	73

Table 8. QWK of the fp32 and INT8 ONNX models (seed 0). The change is fp32 minus INT8, so a positive value is a loss. Seed 0 only, so values differ slightly from Table 3.

Model	Trained on	In dataset fp32	In dataset INT8	Change	Cross dataset fp32	Cross dataset INT8	Change
MobileNetV3	APTOS	0.869	0.391	+0.478	0.312	0.182	+0.130
MobileNetV3	EyePACS	0.444	0.174	+0.270	0.545	0.371	+0.174
EfficientNet Lite0	APTOS	0.863	0.843	+0.020	0.166	0.178	-0.012
EfficientNet Lite0	EyePACS	0.422	0.374	+0.048	0.566	0.542	+0.024
ShuffleNetV2	APTOS	0.857	0.850	+0.007	0.300	0.331	-0.031
ShuffleNetV2	EyePACS	0.426	0.394	+0.032	0.433	0.473	-0.040
GhostNet	APTOS	0.873	0.636	+0.237	0.231	0.219	+0.012
GhostNet	EyePACS	0.451	0.234	+0.217	0.691	0.247	+0.444
MobileViT XS	APTOS	0.891	0.493	+0.398	0.312	0.094	+0.218
MobileViT XS	EyePACS	0.542	0.115	+0.427	0.607	0.063	+0.544
ResNet50	APTOS	0.884	0.888	-0.004	0.143	0.210	-0.067
ResNet50	EyePACS	0.480	0.434	+0.046	0.555	0.419	+0.136

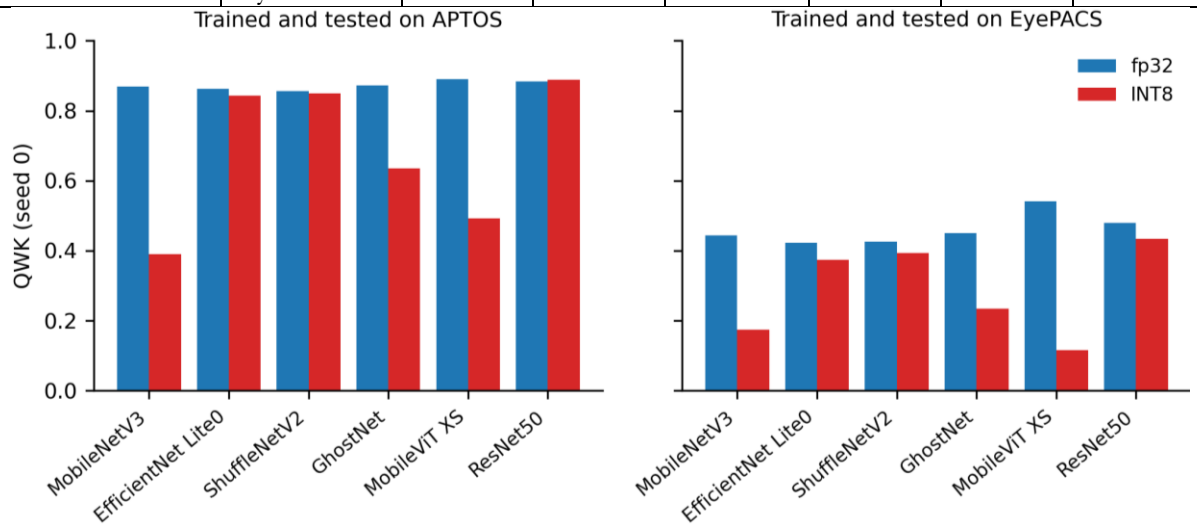


Figure 4. In dataset QWK of the fp32 and INT8 ONNX models for seed 0.

Table 9. Same dataset QWK of fp32 and INT8 models (seed 0) with the default and the best calibration option.

Model, trained on	fp32	Default INT8	Best option
MobileViT XS, APTOS	0.891	0.493	0.595 (percentile)
MobileViT XS, EyePACS	0.542	0.115	0.433 (percentile)
MobileNetV3, APTOS	0.869	0.391	0.510 (MinMax, 50 images)
MobileNetV3, EyePACS	0.444	0.174	0.265 (percentile)
GhostNet, APTOS	0.873	0.636	0.775 (percentile)
GhostNet, EyePACS	0.451	0.234	0.239 (percentile, convolutions only)

The best option is chosen on the test split, so these values are an upper bound.

Grad CAM:

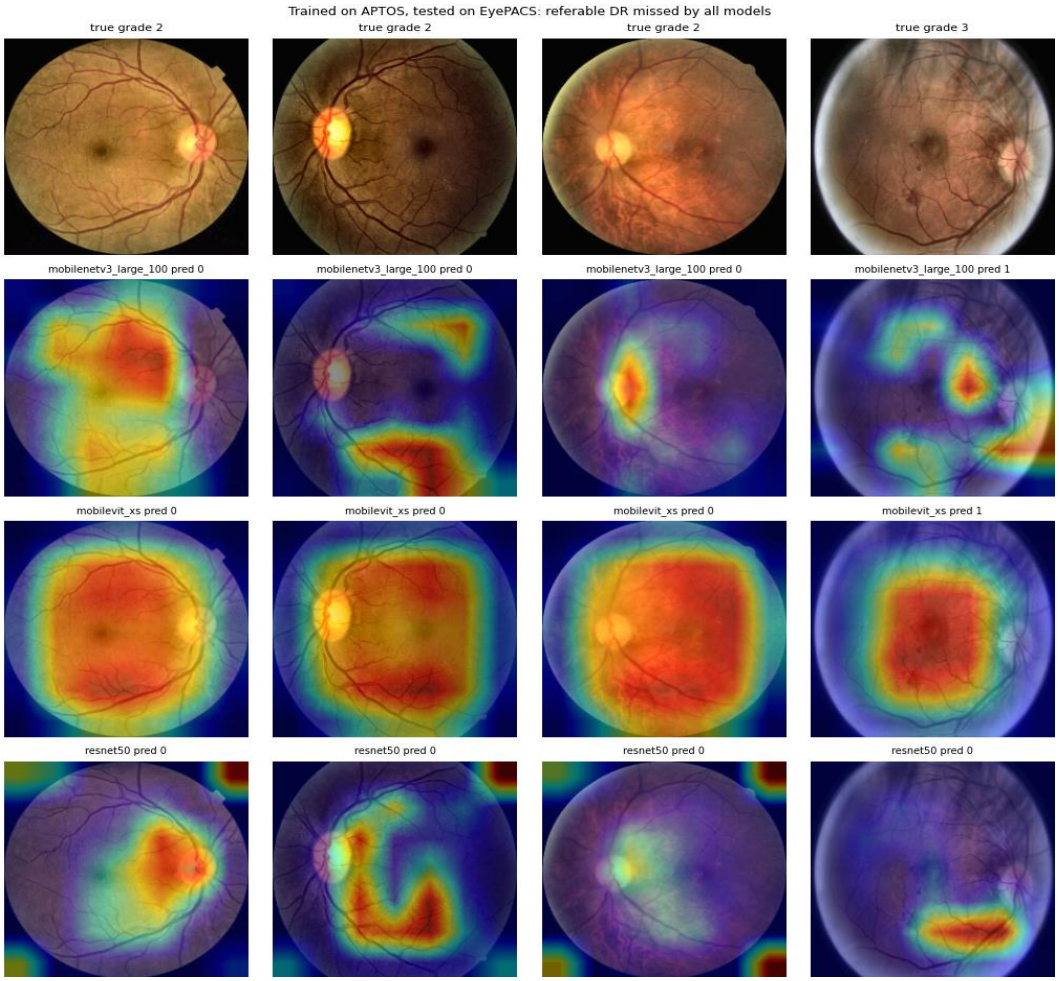


Figure 5. Grad CAM maps for EyePACS test images with referable DR that are missed by all three models trained on APTOS. The top row shows the image and its true grade. The other rows show the predicted grade and the map of each model.

Figure 5 shows Grad CAM maps for four EyePACS test images with referable DR. All three models are trained on APTOS and graded these images as no DR or mild DR. The maps are coarse because the last feature map is small. In several maps the strongest response lies on the image border or the optic disc and not on a visible lesion. ResNet50 also responds in the corners outside the retina in three of the four images. Lesions are hard to see in these four images at this size. Their color and light also differ from APTOS images. Figure 6 shows APTOS test images that all models graded correctly. Hard exudates and bright spots are visible in them. ResNet50 responds mostly over the macular region in three of the four images.

MobileNetV3 and MobileViT XS are less consistent. MobileNetV3 also responds near the image edge in two of the images. For referable DR the expected signs are hemorrhages, microaneurysms, hard exudates, cotton wool spots and, at grade 4, new vessels. Grad CAM shows where the model looked and not why. The maps are coarse, so they cannot outline small lesions. A strong response on the optic disc or image border suggests the models may use cues that are not lesions. The maps are illustrative only. They come from four images each and one seed and there are no lesion masks. They do not show which image features the models use, which fits the caution that explanation methods need validation before they are trusted [15].

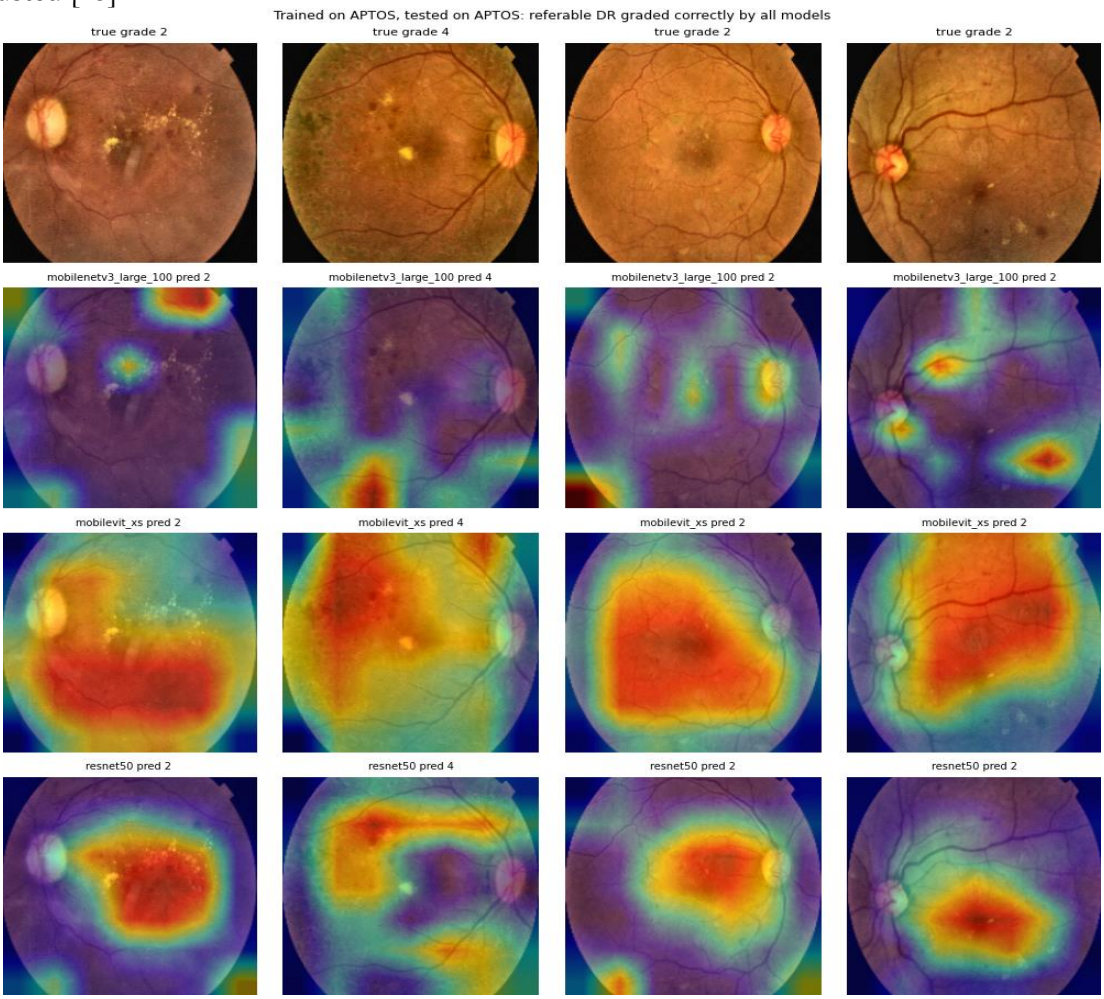


Figure 6. Grad CAM maps for APTOS test images with referable DR that are graded correctly by all three models trained on APTOS.

APPENDIX:

Table 10. QWK with 95 percent bootstrap interval, accuracy, macro F1, and referable DR sensitivity and specificity for every model and setting (means over three seeds).

Model	Setting	QWK (95% interval)	Accuracy	Macro F1	Sens.	Spec.
MobileNetV3	APTOS	0.865 (0.836 to 0.890)	0.811	0.653	0.888	0.945
MobileNetV3	APTOS to EyePACS	0.299 (0.242 to 0.354)	0.689	0.277	0.227	0.958
MobileNetV3	EyePACS	0.465 (0.419 to 0.510)	0.623	0.399	0.616	0.837
MobileNetV3	EyePACS to APTOS	0.606 (0.557 to 0.644)	0.471	0.372	0.896	0.646
EfficientNet Lite0	APTOS	0.867 (0.842 to 0.891)	0.810	0.667	0.860	0.956
EfficientNet Lite0	APTOS to EyePACS	0.203 (0.144 to 0.259)	0.677	0.250	0.164	0.957
EfficientNet Lite0	EyePACS	0.420 (0.375 to 0.465)	0.553	0.401	0.597	0.809
EfficientNet Lite0	EyePACS to APTOS	0.562 (0.521 to 0.601)	0.415	0.327	0.865	0.631
ShuffleNetV2	APTOS	0.871 (0.845 to 0.897)	0.801	0.645	0.829	0.971
ShuffleNetV2	APTOS to EyePACS	0.275 (0.213 to 0.338)	0.722	0.262	0.163	0.977
ShuffleNetV2	EyePACS	0.449 (0.398 to 0.494)	0.557	0.400	0.664	0.813
ShuffleNetV2	EyePACS to APTOS	0.350 (0.306 to 0.395)	0.189	0.197	0.967	0.356
GhostNet	APTOS	0.883 (0.859 to 0.907)	0.816	0.654	0.906	0.950
GhostNet	APTOS to EyePACS	0.229 (0.163 to 0.293)	0.720	0.257	0.141	0.980
GhostNet	EyePACS	0.443 (0.391 to 0.488)	0.598	0.405	0.626	0.833
GhostNet	EyePACS to APTOS	0.583 (0.541 to 0.625)	0.435	0.331	0.956	0.555
MobileViT XS	APTOS	0.894 (0.870 to 0.918)	0.831	0.689	0.891	0.962
MobileViT XS	APTOS to EyePACS	0.342 (0.275 to 0.403)	0.708	0.278	0.220	0.962
MobileViT XS	EyePACS	0.533 (0.485 to 0.577)	0.561	0.434	0.686	0.889
MobileViT XS	EyePACS to APTOS	0.653 (0.618 to 0.686)	0.326	0.310	0.989	0.718
ResNet50	APTOS	0.895 (0.871 to 0.915)	0.822	0.673	0.896	0.966
ResNet50	APTOS to EyePACS	0.162 (0.102 to 0.222)	0.675	0.239	0.147	0.954
ResNet50	EyePACS	0.489 (0.439 to 0.535)	0.621	0.431	0.642	0.835
ResNet50	EyePACS to APTOS	0.556 (0.516 to 0.591)	0.380	0.289	0.977	0.512

Table 11. Paired bootstrap comparison of the best model in each setting with each other model. The difference is best minus other. Intervals are not corrected for 20 comparisons.

Setting	Best model	Compared with	Difference	95% interval	Significant
APTOS	ResNet50	MobileNetV3	0.030	0.007 to 0.051	Yes
APTOS	ResNet50	EfficientNet Lite0	0.028	0.009 to 0.047	Yes
APTOS	ResNet50	ShuffleNetV2	0.024	0.006 to 0.044	Yes
APTOS	ResNet50	GhostNet	0.011	-0.010 to 0.030	No
APTOS	ResNet50	MobileViT XS	0.001	-0.019 to 0.021	No
APTOS to EyePACS	MobileViT XS	MobileNetV3	0.041	-0.003 to 0.086	No
APTOS to EyePACS	MobileViT XS	EfficientNet Lite0	0.138	0.087 to 0.189	Yes
APTOS to EyePACS	MobileViT XS	ShuffleNetV2	0.067	0.021 to 0.113	Yes
APTOS to EyePACS	MobileViT XS	GhostNet	0.113	0.066 to 0.160	Yes
APTOS to EyePACS	MobileViT XS	ResNet50	0.178	0.127 to 0.232	Yes
EyePACS	MobileViT XS	MobileNetV3	0.067	0.027 to 0.105	Yes
EyePACS	MobileViT XS	EfficientNet Lite0	0.112	0.077 to 0.146	Yes
EyePACS	MobileViT XS	ShuffleNetV2	0.083	0.049 to 0.117	Yes
EyePACS	MobileViT XS	GhostNet	0.089	0.056 to 0.123	Yes
EyePACS	MobileViT XS	ResNet50	0.045	0.013 to 0.073	Yes
EyePACS to APTOS	MobileViT XS	MobileNetV3	0.047	0.011 to 0.086	Yes
EyePACS to APTOS	MobileViT XS	EfficientNet Lite0	0.091	0.049 to 0.133	Yes
EyePACS to APTOS	MobileViT XS	ShuffleNetV2	0.302	0.265 to 0.338	Yes
EyePACS to APTOS	MobileViT XS	GhostNet	0.069	0.032 to 0.108	Yes
EyePACS to APTOS	MobileViT XS	ResNet50	0.097	0.068 to 0.125	Yes

Table 12. In dataset QWK, change against baseline and referable DR sensitivity and specificity for each mitigation method (mean ± SD over three seeds).

Model	Dataset	Method	QWK	Change	Seeds better	Sens.	Spec.
MobileViT XS	APTOS	Baseline	0.894 ± 0.005	-	-	0.891	0.962
MobileViT XS	APTOS	Color jitter	0.899 ± 0.010	+0.005	2 of 3	0.902	0.953
MobileViT XS	APTOS	Color normalization	0.881 ± 0.013	-0.013	0 of 3	0.896	0.949
MobileViT XS	APTOS	Mix 10% target	0.901 ± 0.010	+0.007	2 of 3	0.924	0.951
MobileViT XS	EyePACS	Baseline	0.533 ± 0.015	-	-	0.686	0.889

Model	Dataset	Method	QWK	Change	Seeds better	Sens.	Spec.
MobileViT XS	EyePACS	Color jitter	0.529 ± 0.050	-0.004	1 of 3	0.589	0.905
MobileViT XS	EyePACS	Color normalization	0.473 ± 0.017	-0.060	0 of 3	0.689	0.858
MobileViT XS	EyePACS	Mix 10% target	0.570 ± 0.013	+0.038	3 of 3	0.670	0.888
MobileNetV3	APTOS	Baseline	0.865 ± 0.004	-	-	0.888	0.945
MobileNetV3	APTOS	Color jitter	0.859 ± 0.016	-0.006	1 of 3	0.901	0.934
MobileNetV3	APTOS	Color normalization	0.869 ± 0.013	+0.004	2 of 3	0.904	0.940
MobileNetV3	APTOS	Mix 10% target	0.884 ± 0.007	+0.019	3 of 3	0.901	0.944
MobileNetV3	EyePACS	Baseline	0.465 ± 0.024	-	-	0.616	0.837
MobileNetV3	EyePACS	Color jitter	0.435 ± 0.014	-0.030	0 of 3	0.615	0.815
MobileNetV3	EyePACS	Color normalization	0.448 ± 0.020	-0.017	0 of 3	0.611	0.820
MobileNetV3	EyePACS	Mix 10% target	0.496 ± 0.015	+0.032	3 of 3	0.609	0.860

Discussion:

Dataset shift is the main problem in this study. Models that reached 0.865 to 0.895 in dataset QWK on APTOS fell to 0.162 to 0.342 on EyePACS. This agrees with reports that single source models lose accuracy on other DR datasets [6][7]. The color statistics differ between the datasets, mostly in the green and blue channels. This fits the drop but does not prove its cause. Camera type, population, image size and label practice may also differ [5].

MobileViT XS gave the best QWK in three of the four settings. In APTOS to EyePACS it is tied with MobileNetV3. On APTOS in dataset ResNet50 had the highest mean but is not significantly above MobileViT XS or GhostNet. The larger reference model did not generalize better than the small models. It had the lowest QWK from APTOS to EyePACS and the second largest mean same test gap. MobileNetV3 and MobileViT XS had nearly equal mean same test gaps (0.213 and 0.216). MobileNetV3 needs 0.43 GFLOPs against 1.41 GFLOPs for MobileViT XS. Esengönül et al. found EfficientNet based models best on APTOS among five lightweight architectures [10]. Here EfficientNet Lite0 had the second lowest APTOS in dataset score (0.867). The models, preprocessing and training differ, so the two results are not directly comparable.

Mixing a small share of labelled target images is the only strong remedy. It raised the cross dataset QWK of both tested models in both directions. It needs labelled images from the target site. The equal size and extra source controls show that the gain comes from the target images and not from extra training data (Table 6). The controls cover two models and three seeds. Color jitter helped a little in one direction. Color normalization hurt in the other direction and the reason is not tested. Methods built for domain generalization [5][6][7] are not compared here.

The quantization results show that accuracy, generalization and robustness to INT8 are separate properties. The best generalizing model, MobileViT XS, lost most of its QWK with the INT8 method used here. MobileNetV3 and GhostNet are also hurt. ShuffleNetV2, EfficientNet Lite0 and ResNet50 kept their QWK. Attention layers are known to be a poor fit for quantization methods built for convolutional networks [13][30]. That may relate to the MobileViT XS result, but this study did not test the cause. Percentile calibration reduced the loss but did not remove it (Table 9). This suggests that large activation values add to the loss, but we did not test the cause. Quantization aware training adds simulated quantization during fine tuning so the weights can adapt to it [12]. It often recovers much of the loss but needs a training pipeline. We did not test it. Keeping sensitive layers in fp32 is another option. Both are left for future work. On this CPU INT8 also gave no speed gain for most of the lightweight models. Results on ARM hardware or on a phone may differ and need a real device.

For deployment the INT8 format suits ShuffleNetV2 and EfficientNet Lite0 in this study. The fp32 format suits MobileNetV3, GhostNet and MobileViT XS. MobileNetV3 in fp32 has a mean same test gap within 0.003 of MobileViT XS and a median latency of 10.0 ms. MobileViT XS offers the best QWK at 38.5 ms. No baseline model is fit for cross dataset screening. From APTOS to EyePACS the best sensitivity for referable DR is 0.227.

Limitations:

Both datasets are resized copies made by other Kaggle users and their processing is not verified. Images are 224 by 224 pixels and only about 28.5 percent of EyePACS patients are used. The EyePACS in dataset QWK is moderate (0.420 to 0.533). Image size and training set size may be the cause, but this is not tested. Conclusions are therefore relative between models and not an absolute measure of what each model can reach. APTOS has no patient identifiers, so its split is by image. Only exact duplicates are removed, and 158 APTOS images are dropped in total. Three seeds are used. The confidence intervals resample test images and not seeds. No correction is made for multiple comparisons. GPU runs are not bit exact. Only two models are tested for mitigation. Mixing uses labelled target images. The size controls

cover two models and three seeds. The extra source control covers EyePACS only. Latency and memory come from a 4 core Kaggle CPU and not from edge hardware. Edge devices differ in processor, memory bandwidth and power limits, so these values are a proxy. Repeated runs of the same file differed by about 10 percent, so small latency differences are not reliable. INT8 results use one seed and one calibration draw per setting. Calibration options are tested for three models only. Quantization aware training is not tested. Mitigated models are not quantized. Power use is not measured. The Grad CAM maps are qualitative and coarse. They use four images per figure and one seed. No lesion masks were used, so lesion localization was not tested. A quantitative test needs a dataset with lesion masks, such as IDRiD [31]. The models have not been validated using external clinical data and are not for clinical use. EyePACS to APTOS mixing results may be slightly inflated by possible patient overlap in APTOS. The two directions show similar gains (+0.165 and +0.170 for MobileViT XS), which suggests the effect is small.

Conclusion:

This study compared five lightweight networks and a ResNet50 reference for DR grading on APTOS and EyePACS. For objective 1, same dataset QWK is 0.865 to 0.895 on APTOS and 0.420 to 0.533 on EyePACS, and ResNet50 gave no clear gain. For objective 2, cross dataset QWK fell to 0.162 to 0.342 from APTOS to EyePACS and 0.350 to 0.653 the other way. Referable DR sensitivity is only 0.141 to 0.227 from APTOS to EyePACS. For objective 3, mixing about 10 percent of labelled target images is the only fix that worked, and the controls show that the gain comes from target images. For objective 4, INT8 made files 2.8 to 3.9 times smaller. It kept QWK for ShuffleNetV2, EfficientNet Lite0 and ResNet50 but cost 0.217 to 0.478 for MobileNetV3, GhostNet and MobileViT XS. Percentile calibration reduced but did not remove the loss. For objective 5, Grad CAM maps of missed cases often fell on the optic disc or image border. These maps are qualitative only. Three steps are needed before screening use. First, clinical validation on external data from the target site. Second, domain generalization methods with a small set of local labelled images. Third, tests on real edge hardware with power measurement. Quantization aware training or mixed precision should be tested for the three models that lost QWK.

Acknowledgement: The authors thank the Department of Data Science & AI, Khwaja Fareed University of Engineering and Information Technology (KFUEIT), for providing the resources and support needed to carry out this research.

Author's Contribution: Muhammad Adeel Abid is responsible for conceptualization, methodology, and analysis. Naina Noor Sindhu and Faiqa Tanveer are responsible for data preparation and manuscript preparation.

Conflict of Interest: The authors declare that there are no conflicts of interest.

Project Details: Not applicable.

References:

[1] H. Sun *et al.*, "IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045," *Diabetes Res. Clin. Pract.*, vol. 183, Jan. 2022, doi: 10.1016/J.DIABRES.2021.109119.

[2] Z. L. Teo *et al.*, "Global Prevalence of Diabetic Retinopathy and Projection of Burden through 2045: Systematic Review and Meta-analysis," *Ophthalmology*, vol. 128, no. 11, pp. 1580–1591, Nov. 2021, doi: 10.1016/J.OPHTHA.2021.04.027.

[3] I. Saleh *et al.*, "AI-based methods for diagnosing and grading diabetic retinopathy: A comprehensive review," *Artif. Intell. Med.*, vol. 168, Oct. 2025, doi: 10.1016/j.artmed.2025.103221.

[4] H. Guan and M. Liu, "Domain Adaptation for Medical Image Analysis: A Survey," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 3, pp. 1173–1185, Feb. 2021, doi: 10.1109/TBME.2021.3117407.

[5] A. Matsun, D. O. Mohamed, S. Chokuwa, M. Ridzuan, and M. Yaqub, "DGM-DR: Domain Generalization with Mutual Information Regularized Diabetic Retinopathy Classification," Sep. 2023, Accessed: Oct. 20, 2026. [Online]. Available: https://arxiv.org/pdf/2309.09670

[6] H. Che, Y. Cheng, H. Jin, and H. Chen, "Towards Generalizable Diabetic Retinopathy Grading in Unseen Domains," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 14224 LNCS, pp. 430–440, Jul. 2023, doi: 10.1007/978-3-031-43904-9_42.

[7] M. Atwany and M. Yaqub, "DRGen: Domain Generalization in Diabetic Retinopathy Classification," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 13432 LNCS, pp. 635–644, 2022, doi: 10.1007/978-3-031-16434-7_61.

[8] S. Chokuwa and M. H. Khan, "Generalizing Across Domains in Diabetic Retinopathy via Variational Autoencoders," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 14393, pp. 265–274, Sep. 2023, doi: 10.1007/978-3-031-47401-9_26.

[9] J. S. Yoon, K. Oh, Y. Shin, M. A. Mazurowski, and H.-I. Suk, "Domain Generalization for Medical Image Analysis: A Review," *Proc. IEEE*, vol. 112, no. 10, pp. 1583–1609, Dec. 2024, doi: 10.1109/JPROC.2024.3507831.

[10] A. Vinora, J. A. Fathima, and K. J. T. Rakshitaa, "Diabetic Retinopathy Detection using Convolutional Neural Network on Mobile Device," *2025 IEEE 1st Int. Conf. Innov. Eng. Next-Generation Technol. Sustain. ICINVENTS 2025*, 2025, doi: 10.1109/ICINVENTS64613.2025.11401781.

[11] A. Asare, D. A. N. Gookyi, D. Boateng, and F. A. Wulnye, "Deploying and Evaluating Multiple Deep Learning Models on Edge Devices for Diabetic Retinopathy Detection," Jun. 2025, Accessed: Oct. 20, 2026. [Online]. Available: https://arxiv.org/pdf/2506.14834

[12] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer, "A Survey of Quantization Methods for Efficient Neural Network Inference," *Low-Power Comput. Vis.*, pp. 291–326, Mar. 2021, doi: 10.1201/9781003162810-13.

[13] Z. Liu, Y. Wang, K. Han, W. Zhang, S. Ma, and W. Gao, "Post-Training Quantization for Vision Transformer," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 28092–28103, Jun. 2021, Accessed: Oct. 20, 2026. [Online]. Available: https://arxiv.org/pdf/2106.14156

[14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Oct. 2016, doi: 10.1007/s11263-019-01228-7.

[15] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *Lancet Digit. Heal.*, vol. 3, no. 11, pp. e745–e750, Nov. 2021, doi: 10.1016/S2589-7500(21)00208-9.

[16] S. Dey *et al.*, "Managing Diabetic Retinopathy with Deep Learning: A Data Centric Overview," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 16, no. 3, Apr. 2026, doi: 10.1002/widm.70113.

[17] "APTOS 2019 Blindness Detection | Kaggle." Accessed: Oct. 20, 2026. [Online]. Available: https://www.kaggle.com/c/aptos2019-blindness-detection

[18] "Diabetic Retinopathy Detection | Kaggle." Accessed: Oct. 20, 2026. [Online]. Available: https://www.kaggle.com/competitions/diabetic-retinopathy-detection/data

[19] "Diabetic Retinopathy 224x224 (2019 Data)." Accessed: Oct. 20, 2026. [Online].

Available: https://www.kaggle.com/datasets/sovitrath/diabetic-retinopathy-224x224-2019-data

[20] "Diabetic Retinopathy (resized)." Accessed: Oct. 20, 2026. [Online]. Available: https://www.kaggle.com/datasets/tanlikesmath/diabetic-retinopathy-resized/code

[21] S. M. Pizer *et al.*, "Adaptive histogram equalization and its variations," *Comput. vision, Graph. image Process.*, vol. 39, no. 3, pp. 355–368, Sep. 1987, doi: 10.1016/S0734-189X(87)80186-X.

[22] A. Howard *et al.*, "Searching for MobileNetV3," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2019-October, pp. 1314–1324, May 2019, doi: 10.1109/ICCV.2019.00140.

[23] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *Int. Conf. Mach. Learn.*, 2019.

[24] N. Ma, X. Zhang, H. T. Zheng, and J. Sun, "ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11218 LNCS, pp. 122–138, Jul. 2018, doi: 10.1007/978-3-030-01264-9_8.

[25] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More Features from Cheap Operations," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 1577–1586, Nov. 2019, doi: 10.1109/CVPR42600.2020.00165.

[26] M. R. Sachin Mehta, "MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer," *arXiv:2110.02178*, 2021, doi: https://doi.org/10.48550/arXiv.2110.02178.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-December, pp. 770–778, Dec. 2016, doi: 10.1109/CVPR.2016.90.

[28] I. Gulrajani and D. Lopez-Paz, "In Search of Lost Domain Generalization," *ICLR 2021 - 9th Int. Conf. Learn. Represent.*, Jul. 2020, Accessed: Oct. 20, 2026. [Online]. Available: https://arxiv.org/pdf/2007.01434

[29] M. Nagel, M. Fournarakis, R. A. Amjad, Y. Bondarenko, M. van Baalen, and T. Blankevoort, "A White Paper on Neural Network Quantization," Jun. 2021, Accessed: Oct. 20, 2026. [Online]. Available: https://arxiv.org/pdf/2106.08295

[30] Z. Yuan, C. Xue, Y. Chen, Q. Wu, and G. Sun, "PTQ4ViT: Post-training quantization for vision transformers with twin uniform quantization," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 13672 LNCS, pp. 191–207, Nov. 2021, doi: 10.1007/978-3-031-19775-8_12.

[31] P. Porwal *et al.*, "IDRiD: Diabetic Retinopathy – Segmentation and Grading Challenge," *Med. Image Anal.*, vol. 59, Jan. 2020, doi: 10.1016/j.media.2019.101561.



Copyright © by authors and 50Sea. This work is licensed under Creative Commons Attribution 4.0 International License.