Online Resource 1: Supplementary methods and results

Retinal foundation-model versus ImageNet pretraining for deep learning grading of diabetic retinopathy: an external validation study

Pravinraj Ethiraj — Department of Artificial Intelligence and Data Science, Chennai Institute of Technology, Chennai, Tamil Nadu, India — pravinraje.aids2024@citchennai.net

International Ophthalmology

Contents

S0 Changes to the analysis plan **2**

S1 Audit of the third-party IDRiD copy against the official release **3**

S2 Duplicate and leakage checks **4**

S3 Checkpoints, selection rule and the interrupted training run **5**

S4 Messidor-2 examinations **6**

S5 Performance of every model and training run **7**

S6 Primary comparison: run-to-run variation, interval methods and the full internal set **9**

S7 Matched-recipe controls **11**

S8 Case mix and the no-DR/mild boundary **12**

S9 Uncertainty, error detection and referral **15**

S10 Saliency localization against expert lesion masks **17**

S11 Earlier sensitivity analyses (run with the third-party IDRiD copy) **20**

This document reports every analysis performed for the study that is not shown in full in the article. All numbers in sections S4–S10 are generated by script from the per-image predictions in Online Resource 3, the same source as the article's tables and text; the scripts are in Online Resource 4.

Conventions. Unless stated otherwise, intervals are 95% percentile bootstrap intervals (2,000 replicates, seed 42), unadjusted for multiplicity; p values for the four RETFound contrasts are Holm-adjusted. Messidor-2 intervals resample examinations (S4). 'Internal' is the deduplicated APTOS internal test set (n = 692). 'IDRiD' is the official 516-image Disease-Grading release. Section S11 contains analyses run before the official IDRiD release was obtained; they used a 455-image third-party copy of IDRiD (audited in S1), so their IDRiD results are superseded, whereas their APTOS and Messidor-2 results are unaffected.

Abbreviations. AUC, area under the receiver operating characteristic curve; CI, confidence interval; ECE, expected calibration error; LP, linear probe; MCC, Matthews correlation coefficient; MSP, maximum softmax probability; NPV/PPV, negative/positive predictive value; QWK, quadratic weighted kappa.

S0 Changes to the analysis plan

#	Original plan	Change	When and why	Where the original result is reported
1	Third-party copy of IDRiD (455 images)	Official IDRiD release (516 images)	After the label audit in S1 showed missing images and changed labels	S11
2	All 733 internal-test images	41 images with a byte-identical duplicate in training/validation excluded (n = 692)	Duplicates inflate internal agreement and therefore the drop	S6 (second table)
3	Image-level bootstrap; worst-case design-effect bound for Messidor-2	Examination-level bootstrap using the official ADCIS examination file	Fellow-eye errors are correlated (S4)	S11.1
4	One training run per model (seed 42; prespecified primary analysis, retained)	Added exploratory (post hoc) analysis: mean over runs (RETFound full: 3; EfficientNet-B3 and LP: 4), two-level bootstrap	A second RETFound fine-tuning run (seed 43), trained to estimate run-to-run variation, gave a materially larger excess drop than the first; the change was made after that result was known. The extra EfficientNet-B3 and LP runs had been completed earlier as a sensitivity analysis; the third RETFound run (seed 44) was specified before it was trained.	Both analyses: Table 2 of the article; S6
5	—	Crossed-recipe control (EfficientNet-B3 with the RETFound recipe; RETFound with the CNN recipe) added	To separate pretraining from fine-tuning recipe	S7

S1 Audit of the third-party IDRiD copy against the official release

The analysis was initially run on a widely shared third-party copy of IDRiD. Comparison with the official IEEE DataPort release (doi:10.21227/H25W98) gave the following; all main-text IDRiD results use the official release.

- Official images with labels: 516 (train 413, test 103); grade counts 0/1/2/3/4: 168/25/168/93/62; referable 62.6%
- Present in the third-party copy: 455; missing from it: 61
- Image IDs in the copy that are not in the official release: 0
- **Labels in the copy that differ from the official grade: 27** (5.9% of the copy's images), listed below

Image	Official	Copy	Image	Official	Copy	Image	Official	Copy
IDRiD_098	3	4	IDRiD_260	2	1	IDRiD_324	3	4
IDRiD_099	3	4	IDRiD_263	2	0	IDRiD_349	3	2
IDRiD_192	3	4	IDRiD_266	1	0	IDRiD_412	2	1
IDRiD_194	1	0	IDRiD_270	2	0	IDRiD_062test	4	2
IDRiD_198	1	0	IDRiD_278	3	4	IDRiD_063test	1	0
IDRiD_201	2	0	IDRiD_282	2	1	IDRiD_064test	3	1
IDRiD_208	2	0	IDRiD_290	1	0	IDRiD_087test	4	2
IDRiD_215	2	0	IDRiD_299	3	4	IDRiD_088test	2	3
IDRiD_256	1	0	IDRiD_311	2	1	IDRiD_089test	2	3

S2 Duplicate and leakage checks

A byte-level (MD5) comparison of all 5,861 preprocessed images found **no** image shared between APTOS, Messidor-2 and IDRiD. Within APTOS, 41 of the 733 internal-test images have an exact duplicate in the training or validation split (a known property of APTOS 2019). Removing them changes internal QWK as follows (completed RETFound checkpoint): EfficientNet-B3 0.876 → 0.879; ResNet-50 0.875 → 0.880; RETFound linear probe 0.847 → 0.852; RETFound full 0.882 → 0.888 (n = 692). No internal comparison changes. A semantic near-duplicate check in an ImageNet feature space supported the byte-level result:

Feature extractor: ImageNet-pretrained EfficientNet-B3 (1,536-dimensional pooled features, L2-normalized); cosine similarity over all cross-dataset pairs.

Dataset pair	Pairs with cosine ≥ 0.90	≥ 0.95	≥ 0.98
APTOS ↔ Messidor-2	47791	99	0
APTOS ↔ IDRiD	303	0	0
Messidor-2 ↔ IDRiD	507	0	0

APTOS internal-test images with a near-duplicate elsewhere

Comparison pool	Internal-test images with cosine ≥ 0.95	Cosine ≥ 0.90
APTOS train/val	89 / 733	367 / 733
Messidor-2 (external)	7 / 733	214 / 733
IDRiD (external)	0 / 733	38 / 733

S3 Checkpoints, selection rule and the interrupted training run

Every scored checkpoint, the rule used to select it, and its SHA-256 hash (full checkpoint files are available from the author on request). Model code and all training defaults were committed to version control before any external image was scored, and no external result was used for any modeling, checkpoint, threshold or calibration decision.

Model	Selection rule	SHA-256 of checkpoint	Status
EfficientNet-B3 (primary comparator), seed 42	Best APTOS-validation QWK, patience 7	`a595133256a678cb f5ef6daca64842f9 89dc761c61042919 bfe3d3b9926c480b`	Used
ResNet-50 (secondary comparator), seed 42	Best APTOS-validation QWK, patience 7	`225610ffa51b778c 6bfd3abd302aaad9 cd68c7157053e734 31d9aa009a696520`	Used
RETFound linear probe, seed 42	Best APTOS-validation QWK, patience 7	`e7795a1bd2db2095 73e3a868cb927b30 428db265a98f6473 3b5e1ec081fe31c9`	Used
RETFound full fine-tune, seed 42 (completed run)	Best APTOS-validation QWK over the full protocol (≤30 epochs, patience 7): epoch 21 of 28, validation QWK 0.939	`b1d4a7b23b79413c 2d313c4e742c79f7 cf1dc6bfc3a388ef 541868b511fd5933`	Used
RETFound full fine-tune, seed 42 (interrupted run)	Best APTOS-validation QWK at epoch 6 of an interrupted run	`4344dd30356e7ae5 6143c0f6db670bc0 80db9b4d0baeb8fc 866426334990322c`	Superseded (outputs archived)
Pretrained RETFound encoder (color fundus MAE weights)	—	`e1e4f66a1b792eeb 6e2efaf158f33be3 5c8255f36b3d17ed 67cd5129da246485`	Input weights

Interrupted run. The first fully fine-tuned RETFound run was stopped by hardware limits at epoch 6 of the planned 30 (validation QWK 0.935). Scored provisionally, it suggested a significantly larger external drop than EfficientNet-B3 (Messidor-2 QWK 0.368, IDRiD copy 0.720). The pre-specified protocol was then completed (early stopping at epoch 28; best epoch 21; validation QWK 0.939), giving Messidor-2 QWK 0.487, and the conclusion changed from “worse” to “no detectable difference”. Both checkpoints are listed above; all results in the article use the completed run.

Additional runs. RETFound full fine-tuning seeds 43 and 44, EfficientNet-B3 seeds 43–45, RETFound LP seeds 43–45 and the control runs (label smoothing, crossed recipes) used the same selection rule (best APTOS-validation QWK, patience 7); their exact command lines are in Online Resource 4 and their per-image predictions in Online Resource 3.

Table S1. Training configuration

Shared by all models: APTOS split, preprocessing, augmentation, class-weighted cross-entropy, AdamW optimizer, early stopping on APTOS validation QWK (patience 7), mixed precision, 224 × 224 input.

	EfficientNet-B3	ResNet-50	RETFound linear probe	RETFound full fine-tune
Pretraining	ImageNet (supervised)	ImageNet (supervised)	Retinal MAE (Zhou et al., 2023)	Retinal MAE (Zhou et al., 2023)
Trainable layers	All	All	Classification head	All
Learning rate	10^{-4}, cosine	10^{-4}, cosine	10^{-3}, 3-epoch warm-up, cosine	10^{-4}, 3-epoch warm-up, cosine
Weight decay	10^{-4}	10^{-4}	0.05	0.05
Layer-wise decay	–	–	–	0.65
Stochastic depth	0.2[a]	–	0	0.2
Label smoothing	0	0	0.1	0.1
Batch size	32	32	16	16
Maximum epochs	30	30	40	30
Training runs (seeds)	4 (42–45)	1 (42)	4 (42–45)	3 (42–44)
Control runs (seed 42)	Label smoothing 0.1 only; RETFound recipe[b]	–	–	CNN recipe[c]

MAE, masked autoencoder; QWK, quadratic weighted kappa. [a]torchvision default for EfficientNet-B3 (classifier dropout 0.3). [b]Label smoothing 0.1, weight decay 0.05, batch size 16, 3-epoch warm-up. [c]No layer-wise decay, stochastic depth or label smoothing; weight decay 10^{-4}; batch size 16; no warm-up.

S4 Messidor-2 examinations

Messidor-2 contains two macula-centered photographs (one per eye) per examination. Examinations were taken from the official pairing file distributed by ADCIS (one row per examination, listing the left and right image): 874 examinations, 870 with both images gradable and 4 with one gradable image. Fellow eyes had identical reference grades in 76% of examinations and grades within one level in 97%, against 41% exact agreement expected for two random images. The within-examination correlation of EfficientNet-B3 grading errors was 0.55; all Messidor-2 intervals therefore resample whole examinations.

Before the pairing file was available, pairs had been reconstructed from file-name order and optic-disc position. Against the official file, 91% of reconstructed pairs were correct and laterality was inferred correctly for 98% of images. All reported results use the official file.

S5 Performance of every model and training run

Point estimates with 95% bootstrap intervals (2,000 replicates; Messidor-2 resampled by examination). Sensitivity, specificity, PPV and NPV refer to referable DR (grade ≥ 2) at the model's own grade decision, in %. ECE, expected calibration error (15 bins).

APTOS internal (n = 692)

Model	QWK	Accuracy, %	Sensitivity, %	Specificity, %	PPV, %	NPV, %	Referable AUC	ECE
EfficientNet-B3	0.879 (0.850–0.905)	81.1 (78.0–84.0)	89.6 (85.7–93.0)	93.2 (90.6–95.5)	89.2 (85.5–92.8)	93.4 (91.0–95.5)	0.979 (0.970–0.987)	0.094 (0.074–0.123)
ResNet-50	0.880 (0.853–0.904)	80.1 (77.0–82.9)	86.6 (81.9–90.4)	95.5 (93.5–97.4)	92.4 (89.1–95.6)	91.8 (89.1–94.2)	0.978 (0.968–0.986)	0.055 (0.043–0.088)
RETFound LP	0.852 (0.823–0.880)	77.2 (74.0–80.2)	84.7 (80.3–88.6)	95.5 (93.3–97.4)	92.3 (88.8–95.4)	90.8 (88.1–93.2)	0.976 (0.964–0.986)	0.188 (0.166–0.219)
RETFound full	0.888 (0.863–0.910)	79.6 (76.4–82.5)	89.9 (86.1–93.1)	95.5 (93.5–97.4)	92.7 (89.4–95.6)	93.8 (91.4–95.7)	0.964 (0.944–0.981)	0.192 (0.171–0.222)
EfficientNet-B3 + label smoothing (control)	0.874 (0.844–0.900)	78.9 (75.9–81.9)	83.2 (78.7–87.5)	95.3 (93.1–97.2)	91.8 (88.1–95.0)	90.0 (87.1–92.6)	0.937 (0.911–0.960)	0.207 (0.184–0.238)
EfficientNet-B3, RETFound recipe (control)	0.865 (0.833–0.892)	80.1 (77.0–82.9)	86.2 (81.8–90.0)	94.6 (92.3–96.7)	90.9 (87.4–94.4)	91.6 (88.9–94.1)	0.962 (0.944–0.978)	0.173 (0.153–0.206)
RETFound full, CNN recipe (control)	0.888 (0.864–0.909)	77.7 (74.4–80.6)	89.2 (85.1–92.6)	93.6 (91.2–95.8)	89.8 (86.0–93.4)	93.2 (90.7–95.3)	0.979 (0.970–0.987)	0.049 (0.039–0.084)
EfficientNet-B3, seed 43	0.887 (0.860–0.912)	80.9 (78.0–83.8)	86.2 (81.8–90.2)	97.4 (95.8–98.8)	95.5 (92.8–97.9)	91.8 (89.2–94.2)	0.986 (0.979–0.991)	0.081 (0.062–0.110)
RETFound LP, seed 43	0.869 (0.840–0.894)	79.5 (76.4–82.4)	86.9 (82.8–90.5)	95.5 (93.5–97.4)	92.5 (89.0–95.6)	92.0 (89.5–94.3)	0.976 (0.965–0.986)	0.230 (0.204–0.258)
RETFound full, seed 43	0.897 (0.874–0.919)	81.2 (78.2–84.1)	92.9 (89.6–95.8)	92.5 (89.9–94.9)	88.6 (84.8–92.2)	95.4 (93.2–97.3)	0.978 (0.965–0.988)	0.193 (0.171–0.222)
EfficientNet-B3, seed 44	0.881 (0.854–0.907)	81.9 (79.0–84.8)	89.9 (86.2–93.2)	94.1 (91.7–96.2)	90.6 (87.0–93.9)	93.7 (91.2–95.9)	0.976 (0.967–0.985)	0.093 (0.072–0.121)
RETFound LP, seed 44	0.859 (0.829–0.886)	76.7 (73.6–79.8)	80.2 (75.4–84.8)	97.2 (95.4–98.6)	94.7 (91.6–97.4)	88.6 (85.7–91.4)	0.973 (0.959–0.984)	0.224 (0.197–0.253)
RETFound full, seed 44	0.878 (0.851–0.904)	80.5 (77.6–83.4)	88.8 (84.9–92.5)	96.2 (94.3–97.9)	93.7 (90.6–96.4)	93.2 (90.8–95.4)	0.971 (0.955–0.985)	0.193 (0.169–0.221)
EfficientNet-B3, seed 45	0.880 (0.851–0.905)	79.8 (76.6–82.7)	87.7 (83.7–91.4)	93.9 (91.5–96.2)	90.0 (86.3–93.6)	92.3 (89.7–94.7)	0.980 (0.971–0.987)	0.115 (0.094–0.146)
RETFound LP, seed 45	0.865 (0.835–0.892)	77.9 (74.7–80.9)	83.2 (78.5–87.3)	95.3 (93.2–97.2)	91.8 (88.0–95.0)	90.0 (87.2–92.6)	0.975 (0.962–0.985)	0.219 (0.195–0.249)

APTOS internal incl. duplicates (n = 733)

Model	QWK	Accuracy, %	Sensitivity, %	Specificity, %	PPV, %	NPV, %	Referable AUC	ECE
EfficientNet-B3	0.876 (0.848–0.902)	80.4 (77.5–83.1)	90.3 (86.7–93.6)	93.1 (90.7–95.3)	90.0 (86.5–93.2)	93.3 (90.8–95.6)	0.980 (0.971–0.987)	0.101 (0.080–0.130)
ResNet-50	0.875 (0.848–0.899)	78.9 (76.0–81.7)	86.9 (82.9–90.6)	95.4 (93.4–97.3)	92.8 (89.7–95.7)	91.4 (88.7–93.9)	0.978 (0.968–0.986)	0.058 (0.044–0.091)
RETFound LP	0.847 (0.818–0.873)	75.9 (72.7–78.9)	85.9 (81.8–89.8)	95.2 (93.1–97.1)	92.4 (89.1–95.4)	90.8 (88.0–93.3)	0.976 (0.966–0.986)	0.182 (0.159–0.212)
RETFound full	0.882 (0.856–0.906)	78.3 (75.2–81.3)	90.9 (87.4–94.1)	95.2 (93.1–97.1)	92.8 (89.7–95.7)	93.9 (91.5–96.0)	0.967 (0.949–0.983)	0.179 (0.161–0.212)
EfficientNet-B3 + label smoothing (control)	0.867 (0.837–0.895)	77.9 (74.8–80.9)	83.9 (79.6–88.1)	94.9 (92.7–97.0)	91.9 (88.4–95.1)	89.6 (86.7–92.3)	0.939 (0.914–0.962)	0.196 (0.177–0.227)
EfficientNet-B3, RETFound recipe (control)	0.862 (0.831–0.891)	79.3 (76.3–82.1)	86.6 (82.3–90.4)	94.5 (92.3–96.6)	91.5 (88.2–94.7)	91.1 (88.3–93.7)	0.963 (0.946–0.978)	0.166 (0.143–0.194)
RETFound full, CNN recipe (control)	0.881 (0.856–0.903)	76.8 (73.5–79.8)	89.9 (86.3–93.3)	93.1 (90.6–95.4)	89.9 (86.4–93.3)	93.1 (90.4–95.3)	0.978 (0.968–0.985)	0.052 (0.042–0.087)
EfficientNet-B3, seed 43	0.883 (0.854–0.907)	80.1 (77.1–82.9)	86.9 (82.8–90.7)	97.5 (95.9–98.9)	95.9 (93.4–98.1)	91.6 (88.9–94.1)	0.987 (0.980–0.992)	0.091 (0.069–0.118)
RETFound LP, seed 43	0.865 (0.837–0.890)	78.3 (75.2–81.3)	87.9 (84.2–91.5)	95.2 (93.1–97.1)	92.6 (89.3–95.5)	92.0 (89.3–94.4)	0.976 (0.965–0.985)	0.223 (0.196–0.253)
RETFound full, seed 43	0.894 (0.871–0.915)	80.4 (77.4–83.2)	93.6 (90.5–96.3)	92.2 (89.5–94.5)	89.1 (85.6–92.5)	95.5 (93.3–97.3)	0.979 (0.968–0.988)	0.183 (0.163–0.213)
EfficientNet-B3, seed 44	0.880 (0.852–0.904)	81.3 (78.3–84.2)	90.6 (87.0–93.8)	94.0 (91.7–96.3)	91.2 (87.8–94.4)	93.6 (91.1–95.9)	0.978 (0.968–0.986)	0.099 (0.078–0.127)
RETFound LP, seed 44	0.852 (0.822–0.878)	75.3 (72.3–78.4)	81.9 (77.3–86.1)	96.8 (95.1–98.4)	94.6 (91.6–97.3)	88.6 (85.6–91.3)	0.974 (0.961–0.984)	0.213 (0.188–0.244)
RETFound full, seed 44	0.874 (0.847–0.899)	79.4 (76.4–82.1)	89.9 (86.4–93.3)	95.9 (93.9–97.7)	93.7 (90.7–96.5)	93.3 (90.8–95.6)	0.973 (0.958–0.986)	0.188 (0.162–0.215)
EfficientNet-B3, seed 45	0.879 (0.850–0.902)	79.3 (76.3–82.1)	88.6 (84.9–92.1)	93.8 (91.4–95.9)	90.7 (87.0–93.9)	92.3 (89.7–94.7)	0.981 (0.973–0.988)	0.119 (0.100–0.150)
RETFound LP, seed 45	0.858 (0.829–0.884)	76.8 (73.8–79.8)	84.6 (80.1–88.6)	94.9 (92.8–96.9)	92.0 (88.5–95.0)	90.0 (87.0–92.6)	0.976 (0.964–0.985)	0.215 (0.190–0.243)

Model	QWK	Accuracy, %	Sensitivity, %	Specificity, %	PPV, %	NPV, %	Referable AUC	ECE

Messidor-2 (n = 1,744)

Model	QWK	Accuracy, %	Sensitivity, %	Specificity, %	PPV, %	NPV, %	Referable AUC	ECE
EfficientNet-B3	0.499 (0.434–0.557)	62.6 (59.7–65.3)	38.7 (33.3–44.0)	97.4 (96.5–98.3)	84.3 (78.4–89.7)	81.7 (79.4–84.0)	0.842 (0.816–0.867)	0.279 (0.255–0.307)
ResNet-50	0.435 (0.369–0.495)	61.4 (58.5–64.3)	35.7 (30.4–41.3)	95.6 (94.1–96.8)	74.1 (66.7–81.0)	80.7 (78.3–83.1)	0.789 (0.758–0.818)	0.264 (0.239–0.293)
RETFound LP	0.394 (0.321–0.460)	58.9 (56.1–61.8)	42.9 (37.1–48.7)	90.2 (88.3–92.1)	60.9 (53.9–67.5)	81.6 (79.2–84.0)	0.781 (0.750–0.812)	0.121 (0.096–0.149)
RETFound full	0.487 (0.425–0.546)	62.4 (59.7–65.2)	39.2 (34.0–44.7)	97.5 (96.5–98.4)	84.8 (78.9–90.1)	81.9 (79.5–84.1)	0.817 (0.787–0.845)	0.095 (0.072–0.123)
EfficientNet-B3 + label smoothing (control)	0.464 (0.393–0.526)	61.8 (58.9–64.5)	34.4 (29.1–39.8)	98.1 (97.1–98.8)	86.3 (80.0–91.3)	80.8 (78.4–83.1)	0.793 (0.763–0.823)	0.041 (0.025–0.071)
EfficientNet-B3, RETFound recipe (control)	0.485 (0.416–0.545)	61.9 (59.1–64.6)	37.2 (31.9–42.5)	97.2 (96.1–98.2)	82.5 (75.8–88.6)	81.3 (79.0–83.7)	0.810 (0.779–0.838)	0.048 (0.028–0.076)
RETFound full, CNN recipe (control)	0.419 (0.354–0.480)	57.6 (54.8–60.4)	43.8 (38.3–49.0)	89.5 (87.6–91.4)	59.7 (53.2–66.0)	81.8 (79.3–84.1)	0.715 (0.678–0.748)	0.245 (0.221–0.275)
EfficientNet-B3, seed 43	0.469 (0.401–0.528)	61.5 (58.7–64.2)	35.0 (29.6–40.7)	97.9 (97.0–98.7)	85.6 (79.3–90.8)	80.9 (78.5–83.3)	0.843 (0.817–0.868)	0.292 (0.267–0.321)
RETFound LP, seed 43	0.403 (0.330–0.469)	54.6 (51.8–57.4)	39.8 (34.3–45.4)	91.5 (89.6–93.3)	62.5 (55.3–69.3)	81.1 (78.6–83.4)	0.780 (0.750–0.810)	0.132 (0.107–0.159)
RETFound full, seed 43	0.428 (0.357–0.491)	61.5 (58.8–64.4)	31.3 (26.2–36.8)	98.4 (97.5–99.1)	87.7 (80.8–93.3)	80.1 (77.7–82.5)	0.829 (0.802–0.854)	0.091 (0.071–0.119)
EfficientNet-B3, seed 44	0.527 (0.463–0.584)	62.3 (59.5–65.1)	44.4 (39.2–49.9)	96.0 (94.7–97.2)	79.6 (73.7–85.3)	82.9 (80.6–85.1)	0.847 (0.819–0.872)	0.264 (0.240–0.292)
RETFound LP, seed 44	0.408 (0.339–0.474)	57.0 (54.1–59.8)	43.8 (38.1–49.6)	89.7 (87.6–91.6)	60.1 (53.4–66.7)	81.8 (79.3–84.2)	0.789 (0.759–0.818)	0.151 (0.126–0.178)
RETFound full, seed 44	0.401 (0.329–0.466)	61.1 (58.3–64.0)	27.6 (22.5–32.7)	98.9 (98.2–99.5)	90.0 (83.5–95.0)	79.4 (76.9–81.8)	0.799 (0.769–0.828)	0.099 (0.076–0.128)
EfficientNet-B3, seed 45	0.507 (0.440–0.566)	61.8 (59.1–64.6)	39.6 (34.1–45.0)	97.1 (96.0–98.1)	83.0 (77.0–88.7)	81.9 (79.6–84.2)	0.855 (0.831–0.878)	0.290 (0.265–0.317)
RETFound LP, seed 45	0.374 (0.306–0.439)	57.3 (54.6–60.2)	47.0 (41.5–52.6)	86.9 (84.8–89.1)	56.1 (49.7–62.4)	82.2 (79.7–84.6)	0.778 (0.747–0.808)	0.141 (0.116–0.168)

IDRiD official (n = 516)

Model	QWK	Accuracy, %	Sensitivity, %	Specificity, %	PPV, %	NPV, %	Referable AUC	ECE
EfficientNet-B3	0.726 (0.681–0.768)	46.5 (42.4–51.2)	82.7 (78.5–86.6)	90.7 (86.5–94.7)	93.7 (90.8–96.3)	75.8 (70.2–81.1)	0.927 (0.905–0.948)	0.320 (0.276–0.361)
ResNet-50	0.662 (0.604–0.711)	42.2 (38.2–46.5)	71.5 (66.8–76.5)	95.3 (92.2–98.0)	96.2 (93.7–98.4)	66.7 (61.2–72.2)	0.912 (0.887–0.937)	0.282 (0.244–0.331)
RETFound LP	0.590 (0.533–0.643)	31.2 (27.1–35.3)	82.7 (78.6–86.7)	65.8 (59.2–72.4)	80.2 (75.8–84.4)	69.4 (63.1–75.9)	0.882 (0.853–0.911)	0.154 (0.118–0.194)
RETFound full	0.703 (0.653–0.747)	40.3 (36.0–44.6)	73.4 (68.5–78.3)	92.7 (88.9–96.2)	94.4 (91.3–97.1)	67.5 (62.0–73.3)	0.876 (0.846–0.907)	0.170 (0.131–0.215)
EfficientNet-B3 + label smoothing (control)	0.681 (0.629–0.727)	40.1 (36.2–44.6)	74.0 (69.2–78.8)	95.9 (92.7–98.5)	96.8 (94.4–98.8)	68.8 (63.2–74.5)	0.892 (0.863–0.920)	0.185 (0.150–0.233)
EfficientNet-B3, RETFound recipe (control)	0.690 (0.641–0.733)	41.1 (37.0–45.5)	75.5 (71.0–80.2)	95.3 (92.2–98.0)	96.4 (94.0–98.5)	70.0 (64.5–75.6)	0.900 (0.873–0.926)	0.165 (0.129–0.211)
RETFound full, CNN recipe (control)	0.741 (0.699–0.781)	50.8 (46.5–55.2)	78.9 (74.3–83.4)	92.2 (88.2–95.8)	94.4 (91.6–97.0)	72.4 (66.9–78.2)	0.932 (0.911–0.951)	0.156 (0.127–0.203)
EfficientNet-B3, seed 43	0.737 (0.693–0.777)	47.3 (43.0–51.7)	77.7 (72.9–82.2)	96.9 (94.3–99.0)	97.7 (95.7–99.2)	72.2 (66.7–77.6)	0.933 (0.911–0.953)	0.266 (0.228–0.314)
RETFound LP, seed 43	0.587 (0.535–0.637)	31.6 (27.7–35.7)	78.9 (74.4–83.4)	65.3 (58.8–71.8)	79.2 (74.7–83.4)	64.9 (58.3–71.9)	0.865 (0.835–0.894)	0.136 (0.103–0.179)
RETFound full, seed 43	0.674 (0.624–0.725)	45.7 (41.5–50.2)	83.6 (79.4–87.6)	80.8 (75.3–86.1)	87.9 (84.2–91.5)	74.6 (68.7–80.5)	0.922 (0.899–0.945)	0.046 (0.034–0.096)
EfficientNet-B3, seed 44	0.671 (0.613–0.724)	41.7 (37.6–45.9)	80.8 (76.6–85.0)	92.2 (88.3–95.9)	94.6 (91.8–97.1)	74.2 (68.8–79.6)	0.920 (0.898–0.942)	0.377 (0.340–0.425)
RETFound LP, seed 44	0.592 (0.540–0.643)	29.8 (26.0–33.9)	79.9 (75.4–84.3)	65.3 (58.9–71.5)	79.4 (74.9–83.5)	66.0 (59.4–72.7)	0.873 (0.844–0.902)	0.161 (0.127–0.202)
RETFound full, seed 44	0.709 (0.666–0.749)	41.5 (37.2–45.5)	65.9 (60.8–71.3)	95.3 (92.3–98.0)	95.9 (93.1–98.3)	62.6 (57.1–68.2)	0.904 (0.878–0.929)	0.091 (0.070–0.141)
EfficientNet-B3, seed 45	0.704 (0.656–0.745)	40.7 (36.6–45.0)	77.4 (72.8–82.0)	94.8 (91.5–97.6)	96.2 (93.7–98.2)	71.5 (65.9–77.0)	0.927 (0.905–0.949)	0.395 (0.353–0.437)
RETFound LP, seed 45	0.609 (0.558–0.657)	31.2 (27.3–35.3)	77.7 (73.3–82.4)	66.3 (59.8–72.7)	79.4 (74.9–83.8)	64.0 (57.6–70.6)	0.872 (0.842–0.901)	0.144 (0.111–0.186)

S6 Primary comparison: run-to-run variation, interval methods and the full internal set

Excess drop = (internal − external QWK of the model family, averaged over its runs) − (same for EfficientNet-B3). *Conditional* intervals resample images (examinations for Messidor-2) with the observed runs fixed; *two-level* intervals also resample runs within each family. This multi-run analysis was introduced after the second RETFound run had been observed and is exploratory (post hoc); the prespecified analysis is the seed-42 single-run comparison in the second table of each block. 95% CIs are unadjusted for multiplicity; p values are Holm-adjusted. Holm adjustment over the four RETFound contrasts.

Deduplicated internal set (n = 692)

Family	External	Runs	Drop per run	Mean drop	SD	Excess drop (conditional 95% CI)	Holm p	Two-level 95% CI	Holm p (two-level)
EfficientNet-B3	Messidor-2	4	0.381, 0.418, 0.355, 0.373	0.382	0.027	reference	–	–	–
RETFound LP	Messidor-2	4	0.459, 0.467, 0.451, 0.491	0.467	0.017	+0.085 (+0.028 to +0.139)	0.008	+0.021 to +0.146	0.039
RETFound full	Messidor-2	3	0.401, 0.470, 0.477	0.449	0.042	+0.068 (+0.027 to +0.108)	<0.001	+0.003 to +0.134	0.074
EfficientNet-B3	IDRiD	4	0.153, 0.151, 0.210, 0.176	0.173	0.028	reference	–	–	–
RETFound LP	IDRiD	4	0.263, 0.282, 0.267, 0.256	0.267	0.011	+0.094 (+0.042 to +0.145)	0.003	+0.034 to +0.155	0.016
RETFound full	IDRiD	3	0.186, 0.224, 0.169	0.193	0.028	+0.020 (-0.026 to +0.065)	0.386	-0.041 to +0.080	0.550

Prespecified single-run analysis (seed 42) versus EfficientNet-B3 seed 42, and secondary contrasts:

Model	External	Drop	Excess drop vs EfficientNet-B3 (95% CI)	p	Holm p
ResNet-50	Messidor-2	0.446	+0.065 (+0.014 to +0.116)	0.011	–
ResNet-50	IDRiD	0.218	+0.065 (+0.013 to +0.119)	0.011	–
RETFound LP	Messidor-2	0.459	+0.078 (+0.012 to +0.141)	0.022	0.066
RETFound LP	IDRiD	0.263	+0.110 (+0.046 to +0.175)	0.002	0.008
RETFound full	Messidor-2	0.401	+0.020 (-0.024 to +0.064)	0.355	0.492
RETFound full	IDRiD	0.186	+0.033 (-0.023 to +0.091)	0.246	0.492
EfficientNet-B3 + label smoothing (control)	Messidor-2	0.409	+0.029 (-0.010 to +0.071)	0.161	–
EfficientNet-B3 + label smoothing (control)	IDRiD	0.192	+0.039 (0.000 to +0.080)	0.053	–
EfficientNet-B3, RETFound recipe (control)	Messidor-2	0.380	-0.001 (-0.038 to +0.034)	0.949	–
EfficientNet-B3, RETFound recipe (control)	IDRiD	0.175	+0.022 (-0.025 to +0.068)	0.357	–
RETFound full, seed 43	Messidor-2	0.470	+0.089 (+0.040 to +0.138)	<0.001	–
RETFound full, seed 43	IDRiD	0.224	+0.071 (+0.012 to +0.130)	0.019	–
RETFound full, seed 44	Messidor-2	0.477	+0.097 (+0.046 to +0.147)	<0.001	–
RETFound full, seed 44	IDRiD	0.169	+0.017 (-0.042 to +0.072)	0.570	–
RETFound full, CNN recipe (control)	Messidor-2	0.469	+0.088 (+0.043 to +0.132)	<0.001	–
RETFound full, CNN recipe (control)	IDRiD	0.147	-0.006 (-0.063 to +0.047)	0.801	–
EfficientNet-B3, seed 43	Messidor-2	0.418	+0.037 (+0.006 to +0.071)	0.015	–
EfficientNet-B3, seed 43	IDRiD	0.151	-0.002 (-0.044 to +0.040)	0.940	–
EfficientNet-B3, seed 44	Messidor-2	0.355	-0.026 (-0.064 to +0.012)	0.175	–
EfficientNet-B3, seed 44	IDRiD	0.210	+0.058 (+0.016 to +0.101)	0.003	–
EfficientNet-B3, seed 45	Messidor-2	0.373	-0.007 (-0.042 to +0.029)	0.695	–
EfficientNet-B3, seed 45	IDRiD	0.176	+0.023 (-0.014 to +0.061)	0.247	–
RETFound LP, seed 43	Messidor-2	0.467	+0.086 (+0.023 to +0.148)	0.011	–
RETFound LP, seed 43	IDRiD	0.282	+0.129 (+0.071 to +0.187)	<0.001	–
RETFound LP, seed 44	Messidor-2	0.451	+0.071 (+0.005 to +0.131)	0.035	–
RETFound LP, seed 44	IDRiD	0.267	+0.114 (+0.053 to +0.172)	0.001	–
RETFound LP, seed 45	Messidor-2	0.491	+0.111 (+0.046 to +0.174)	0.001	–
RETFound LP, seed 45	IDRiD	0.256	+0.103 (+0.044 to +0.163)	0.001	–

All 733 internal-test images (original plan)

Family	External	Runs	Drop per run	Mean drop	SD	Excess drop (conditional 95% CI)	Holm p	Two-level 95% CI	Holm p (two-level)
EfficientNet-B3	Messidor-2	4	0.377, 0.413, 0.353, 0.372	0.379	0.025	reference	–	–	–
RETFound LP	Messidor-2	4	0.453, 0.462, 0.444, 0.484	0.461	0.017	+0.082 (+0.024 to +0.137)	0.010	+0.017 to +0.145	0.048
RETFound full	Messidor-2	3	0.395, 0.466, 0.473	0.445	0.043	+0.066 (+0.025 to +0.107)	0.009	+0.005 to +0.130	0.058
EfficientNet-B3	IDRiD	4	0.150, 0.146, 0.209, 0.175	0.170	0.029	reference	–	–	–
RETFound LP	IDRiD	4	0.257, 0.278, 0.260, 0.249	0.261	0.012	+0.091 (+0.037 to +0.144)	<0.001	+0.031 to +0.151	0.024
RETFound full	IDRiD	3	0.180, 0.220, 0.166	0.188	0.028	+0.018 (-0.029 to +0.065)	0.447	-0.045 to +0.078	0.566

Prespecified single-run analysis (seed 42) versus EfficientNet-B3 seed 42, and secondary contrasts:

Model	External	Drop	Excess drop vs EfficientNet-B3 (95% CI)	p	Holm p
ResNet-50	Messidor-2	0.440	+0.063 (+0.011 to +0.113)	0.016	–
ResNet-50	IDRiD	0.212	+0.063 (+0.013 to +0.115)	0.016	–
RETFound LP	Messidor-2	0.453	+0.076 (+0.008 to +0.142)	0.022	0.066
RETFound LP	IDRiD	0.257	+0.108 (+0.042 to +0.175)	0.004	0.016
RETFound full	Messidor-2	0.395	+0.018 (-0.027 to +0.061)	0.457	0.628
RETFound full	IDRiD	0.180	+0.030 (-0.025 to +0.087)	0.314	0.628
EfficientNet-B3 + label smoothing (control)	Messidor-2	0.403	+0.026 (-0.014 to +0.067)	0.207	–
EfficientNet-B3 + label smoothing (control)	IDRiD	0.186	+0.037 (-0.005 to +0.078)	0.086	–
EfficientNet-B3, RETFound recipe (control)	Messidor-2	0.377	0.000 (-0.035 to +0.035)	0.977	–
EfficientNet-B3, RETFound recipe (control)	IDRiD	0.173	+0.023 (-0.022 to +0.068)	0.305	–
RETFound full, seed 43	Messidor-2	0.466	+0.089 (+0.039 to +0.141)	0.001	–
RETFound full, seed 43	IDRiD	0.220	+0.071 (+0.012 to +0.128)	0.020	–
RETFound full, seed 44	Messidor-2	0.473	+0.096 (+0.045 to +0.148)	<0.001	–
RETFound full, seed 44	IDRiD	0.165	+0.016 (-0.042 to +0.072)	0.600	–
RETFound full, CNN recipe (control)	Messidor-2	0.462	+0.085 (+0.038 to +0.129)	<0.001	–
RETFound full, CNN recipe (control)	IDRiD	0.140	-0.010 (-0.064 to +0.045)	0.696	–
EfficientNet-B3, seed 43	Messidor-2	0.413	+0.036 (+0.005 to +0.068)	0.018	–
EfficientNet-B3, seed 43	IDRiD	0.146	-0.003 (-0.046 to +0.037)	0.879	–
EfficientNet-B3, seed 44	Messidor-2	0.353	-0.024 (-0.062 to +0.015)	0.212	–
EfficientNet-B3, seed 44	IDRiD	0.209	+0.059 (+0.018 to +0.104)	0.003	–
EfficientNet-B3, seed 45	Messidor-2	0.372	-0.005 (-0.037 to +0.031)	0.742	–
EfficientNet-B3, seed 45	IDRiD	0.175	+0.025 (-0.013 to +0.061)	0.179	–
RETFound LP, seed 43	Messidor-2	0.462	+0.085 (+0.022 to +0.149)	0.011	–
RETFound LP, seed 43	IDRiD	0.278	+0.128 (+0.069 to +0.188)	<0.001	–
RETFound LP, seed 44	Messidor-2	0.444	+0.067 (+0.003 to +0.129)	0.038	–
RETFound LP, seed 44	IDRiD	0.260	+0.110 (+0.048 to +0.169)	0.001	–
RETFound LP, seed 45	Messidor-2	0.484	+0.107 (+0.041 to +0.169)	0.003	–
RETFound LP, seed 45	IDRiD	0.249	+0.099 (+0.039 to +0.159)	0.002	–

S7 Matched-recipe controls

Crossed-recipe experiment (seed 42): RETFound full fine-tuning with the CNN recipe compared with EfficientNet-B3 (CNN recipe), and RETFound full fine-tuning (RETFound recipe) compared with EfficientNet-B3 trained with the RETFound recipe. Recipes are given in Table S1 (S3).

Comparison	External	Metric	Drop, RETFound	Drop, EfficientNet-B3	Excess drop (95% CI)	p
RETFound vs EfficientNet-B3, both RETFound recipe	Messidor-2	QWK	0.401	0.380	+0.021 (-0.023 to +0.067)	0.353
RETFound vs EfficientNet-B3, both RETFound recipe	Messidor-2	QWK, both corrections	0.124	0.072	+0.052 (-0.018 to +0.125)	0.136
RETFound vs EfficientNet-B3, both RETFound recipe	Messidor-2	Referable-DR AUC	0.147	0.153	-0.005 (-0.035 to +0.025)	0.705
RETFound vs EfficientNet-B3, both RETFound recipe	IDRiD	QWK	0.186	0.175	+0.010 (-0.044 to +0.068)	0.715
RETFound vs EfficientNet-B3, both RETFound recipe	IDRiD	QWK, both corrections	0.120	0.040	+0.080 (+0.001 to +0.160)	0.050
RETFound vs EfficientNet-B3, both RETFound recipe	IDRiD	Referable-DR AUC	0.088	0.063	+0.025 (-0.011 to +0.062)	0.167
RETFound vs EfficientNet-B3, both CNN recipe	Messidor-2	QWK	0.469	0.381	+0.088 (+0.043 to +0.132)	<0.001
RETFound vs EfficientNet-B3, both CNN recipe	Messidor-2	QWK, both corrections	0.199	0.143	+0.057 (-0.012 to +0.124)	0.100
RETFound vs EfficientNet-B3, both CNN recipe	Messidor-2	Referable-DR AUC	0.264	0.137	+0.127 (+0.098 to +0.157)	<0.001
RETFound vs EfficientNet-B3, both CNN recipe	IDRiD	QWK	0.147	0.153	-0.006 (-0.063 to +0.047)	0.801
RETFound vs EfficientNet-B3, both CNN recipe	IDRiD	QWK, both corrections	0.067	0.047	+0.019 (-0.060 to +0.101)	0.632
RETFound vs EfficientNet-B3, both CNN recipe	IDRiD	Referable-DR AUC	0.047	0.052	-0.005 (-0.024 to +0.014)	0.610

S8 Case mix and the no-DR/mild boundary

Internal (n = 692) minus external QWK under each correction, seed-42 runs, and the difference from EfficientNet-B3. These corrections describe how the drop depends on case mix and on the no-DR/mild boundary; they do not identify its cause.

Model	External	Measure	Drop (95% CI)	Excess drop vs EfficientNet-B3
EfficientNet-B3	Messidor-2	QWK	0.381 (0.316–0.451)	–
EfficientNet-B3	Messidor-2	QWK, case mix matched	0.293 (0.235–0.363)	–
EfficientNet-B3	Messidor-2	QWK, grades 0+1 merged	0.203 (0.113–0.297)	–
EfficientNet-B3	Messidor-2	QWK, both corrections	0.143 (0.047–0.249)	–
EfficientNet-B3	IDRiD	QWK	0.153 (0.103–0.205)	–
EfficientNet-B3	IDRiD	QWK, case mix matched	0.186 (0.126–0.250)	–
EfficientNet-B3	IDRiD	QWK, grades 0+1 merged	0.073 (-0.004–0.146)	–
EfficientNet-B3	IDRiD	QWK, both corrections	0.047 (-0.033–0.128)	–
ResNet-50	Messidor-2	QWK	0.446 (0.380–0.515)	+0.065 (+0.014 to +0.116)
ResNet-50	Messidor-2	QWK, case mix matched	0.379 (0.301–0.456)	+0.087 (+0.010 to +0.158)
ResNet-50	Messidor-2	QWK, grades 0+1 merged	0.254 (0.165–0.347)	+0.052 (-0.026 to +0.125)
ResNet-50	Messidor-2	QWK, both corrections	0.215 (0.107–0.319)	+0.072 (-0.036 to +0.167)
ResNet-50	IDRiD	QWK	0.218 (0.163–0.278)	+0.065 (+0.013 to +0.119)
ResNet-50	IDRiD	QWK, case mix matched	0.242 (0.175–0.313)	+0.056 (-0.007 to +0.122)
ResNet-50	IDRiD	QWK, grades 0+1 merged	0.161 (0.081–0.240)	+0.088 (+0.021 to +0.157)
ResNet-50	IDRiD	QWK, both corrections	0.140 (0.054–0.226)	+0.092 (+0.013 to +0.170)
RETFound LP	Messidor-2	QWK	0.459 (0.386–0.535)	+0.078 (+0.012 to +0.141)
RETFound LP	Messidor-2	QWK, case mix matched	0.307 (0.254–0.365)	+0.015 (-0.062 to +0.084)
RETFound LP	Messidor-2	QWK, grades 0+1 merged	0.311 (0.213–0.409)	+0.108 (+0.020 to +0.195)
RETFound LP	Messidor-2	QWK, both corrections	0.137 (0.057–0.215)	-0.006 (-0.112 to +0.093)
RETFound LP	IDRiD	QWK	0.263 (0.199–0.324)	+0.110 (+0.046 to +0.175)
RETFound LP	IDRiD	QWK, case mix matched	0.369 (0.297–0.437)	+0.182 (+0.105 to +0.259)
RETFound LP	IDRiD	QWK, grades 0+1 merged	0.123 (0.044–0.198)	+0.050 (-0.027 to +0.130)
RETFound LP	IDRiD	QWK, both corrections	0.198 (0.113–0.284)	+0.151 (+0.061 to +0.236)
RETFound full	Messidor-2	QWK	0.401 (0.335–0.470)	+0.020 (-0.024 to +0.064)
RETFound full	Messidor-2	QWK, case mix matched	0.289 (0.239–0.339)	-0.004 (-0.066 to +0.045)
RETFound full	Messidor-2	QWK, grades 0+1 merged	0.208 (0.123–0.292)	+0.006 (-0.059 to +0.071)
RETFound full	Messidor-2	QWK, both corrections	0.124 (0.045–0.205)	-0.019 (-0.103 to +0.052)
RETFound full	IDRiD	QWK	0.186 (0.133–0.241)	+0.033 (-0.023 to +0.091)
RETFound full	IDRiD	QWK, case mix matched	0.234 (0.173–0.301)	+0.048 (-0.022 to +0.119)
RETFound full	IDRiD	QWK, grades 0+1 merged	0.121 (0.043–0.194)	+0.048 (-0.023 to +0.127)
RETFound full	IDRiD	QWK, both corrections	0.120 (0.040–0.202)	+0.072 (-0.006 to +0.155)
EfficientNet-B3 + label smoothing (control)	Messidor-2	QWK	0.409 (0.343–0.485)	+0.029 (-0.010 to +0.071)
EfficientNet-B3 + label smoothing (control)	Messidor-2	QWK, case mix matched	0.306 (0.248–0.374)	+0.013 (-0.043 to +0.063)
EfficientNet-B3 + label smoothing (control)	Messidor-2	QWK, grades 0+1 merged	0.215 (0.130–0.312)	+0.013 (-0.049 to +0.081)
EfficientNet-B3 + label smoothing (control)	Messidor-2	QWK, both corrections	0.127 (0.047–0.216)	-0.016 (-0.101 to +0.062)
EfficientNet-B3 + label smoothing (control)	IDRiD	QWK	0.192 (0.136–0.250)	+0.039 (0.000 to +0.080)
EfficientNet-B3 + label smoothing (control)	IDRiD	QWK, case mix matched	0.218 (0.152–0.287)	+0.032 (-0.016 to +0.082)
EfficientNet-B3 + label smoothing (control)	IDRiD	QWK, grades 0+1 merged	0.134 (0.055–0.208)	+0.061 (+0.004 to +0.121)
EfficientNet-B3 + label smoothing (control)	IDRiD	QWK, both corrections	0.108 (0.023–0.191)	+0.060 (0.000 to +0.122)
EfficientNet-B3, RETFound recipe (control)	Messidor-2	QWK	0.380 (0.312–0.455)	-0.001 (-0.038 to +0.034)
EfficientNet-B3, RETFound recipe (control)	Messidor-2	QWK, case mix matched	0.272 (0.214–0.335)	-0.021 (-0.070 to +0.021)
EfficientNet-B3, RETFound recipe (control)	Messidor-2	QWK, grades 0+1 merged	0.166 (0.072–0.265)	-0.037 (-0.098 to +0.019)
EfficientNet-B3, RETFound recipe (control)	Messidor-2	QWK, both corrections	0.072 (-0.015–0.159)	-0.071 (-0.151 to -0.007)
EfficientNet-B3, RETFound recipe (control)	IDRiD	QWK	0.175 (0.120–0.230)	+0.022 (-0.025 to +0.068)
EfficientNet-B3, RETFound recipe (control)	IDRiD	QWK, case mix matched	0.190 (0.131–0.249)	+0.004 (-0.055 to +0.061)
EfficientNet-B3, RETFound recipe (control)	IDRiD	QWK, grades 0+1 merged	0.089 (0.007–0.166)	+0.016 (-0.050 to +0.077)
EfficientNet-B3, RETFound recipe (control)	IDRiD	QWK, both corrections	0.040 (-0.044–0.119)	-0.007 (-0.082 to +0.064)
RETFound full, seed 43	Messidor-2	QWK	0.470 (0.397–0.543)	+0.089 (+0.040 to +0.138)
RETFound full, seed 43	Messidor-2	QWK, case mix matched	0.355 (0.289–0.427)	+0.062 (-0.012 to +0.125)
RETFound full, seed 43	Messidor-2	QWK, grades 0+1 merged	0.288 (0.196–0.380)	+0.085 (+0.012 to +0.157)
RETFound full, seed 43	Messidor-2	QWK, both corrections	0.199 (0.103–0.306)	+0.056 (-0.045 to +0.141)
RETFound full, seed 43	IDRiD	QWK	0.224 (0.167–0.279)	+0.071 (+0.012 to +0.130)
RETFound full, seed 43	IDRiD	QWK, case mix matched	0.296 (0.228–0.361)	+0.110 (+0.035 to +0.178)

Model	External	Measure	Drop (95% CI)	Excess drop vs EfficientNet-B3
RETFound full, seed 43	IDRiD	QWK, grades 0+1 merged	0.187 (0.110–0.258)	+0.114 (+0.039 to +0.195)
RETFound full, seed 43	IDRiD	QWK, both corrections	0.232 (0.147–0.313)	+0.185 (+0.092 to +0.275)
RETFound full, seed 44	Messidor-2	QWK	0.477 (0.405–0.553)	+0.097 (+0.046 to +0.147)
RETFound full, seed 44	Messidor-2	QWK, case mix matched	0.366 (0.293–0.444)	+0.073 (0.000 to +0.137)
RETFound full, seed 44	Messidor-2	QWK, grades 0+1 merged	0.279 (0.185–0.382)	+0.076 (+0.003 to +0.150)
RETFound full, seed 44	Messidor-2	QWK, both corrections	0.192 (0.088–0.305)	+0.049 (-0.050 to +0.135)
RETFound full, seed 44	IDRiD	QWK	0.169 (0.118–0.219)	+0.017 (-0.042 to +0.072)
RETFound full, seed 44	IDRiD	QWK, case mix matched	0.181 (0.131–0.230)	-0.005 (-0.078 to +0.060)
RETFound full, seed 44	IDRiD	QWK, grades 0+1 merged	0.132 (0.056–0.206)	+0.059 (-0.020 to +0.139)
RETFound full, seed 44	IDRiD	QWK, both corrections	0.094 (0.023–0.161)	+0.046 (-0.042 to +0.130)
RETFound full, CNN recipe (control)	Messidor-2	QWK	0.469 (0.403–0.538)	+0.088 (+0.043 to +0.132)
RETFound full, CNN recipe (control)	Messidor-2	QWK, case mix matched	0.355 (0.301–0.412)	+0.063 (+0.021 to +0.105)
RETFound full, CNN recipe (control)	Messidor-2	QWK, grades 0+1 merged	0.279 (0.195–0.365)	+0.077 (+0.016 to +0.136)
RETFound full, CNN recipe (control)	Messidor-2	QWK, both corrections	0.199 (0.116–0.291)	+0.057 (-0.012 to +0.124)
RETFound full, CNN recipe (control)	IDRiD	QWK	0.147 (0.100–0.194)	-0.006 (-0.063 to +0.047)
RETFound full, CNN recipe (control)	IDRiD	QWK, case mix matched	0.162 (0.113–0.211)	-0.025 (-0.092 to +0.043)
RETFound full, CNN recipe (control)	IDRiD	QWK, grades 0+1 merged	0.095 (0.028–0.161)	+0.022 (-0.047 to +0.093)
RETFound full, CNN recipe (control)	IDRiD	QWK, both corrections	0.067 (0.004–0.129)	+0.019 (-0.060 to +0.101)
EfficientNet-B3, seed 43	Messidor-2	QWK	0.418 (0.354–0.492)	+0.037 (+0.006 to +0.071)
EfficientNet-B3, seed 43	Messidor-2	QWK, case mix matched	0.344 (0.278–0.420)	+0.051 (+0.014 to +0.092)
EfficientNet-B3, seed 43	Messidor-2	QWK, grades 0+1 merged	0.233 (0.147–0.328)	+0.030 (-0.018 to +0.081)
EfficientNet-B3, seed 43	Messidor-2	QWK, both corrections	0.191 (0.090–0.301)	+0.049 (-0.008 to +0.104)
EfficientNet-B3, seed 43	IDRiD	QWK	0.151 (0.104–0.201)	-0.002 (-0.044 to +0.040)
EfficientNet-B3, seed 43	IDRiD	QWK, case mix matched	0.153 (0.104–0.204)	-0.034 (-0.083 to +0.015)
EfficientNet-B3, seed 43	IDRiD	QWK, grades 0+1 merged	0.091 (0.020–0.160)	+0.019 (-0.041 to +0.072)
EfficientNet-B3, seed 43	IDRiD	QWK, both corrections	0.047 (-0.022–0.114)	-0.001 (-0.069 to +0.062)
EfficientNet-B3, seed 44	Messidor-2	QWK	0.355 (0.293–0.425)	-0.026 (-0.064 to +0.012)
EfficientNet-B3, seed 44	Messidor-2	QWK, case mix matched	0.250 (0.198–0.307)	-0.043 (-0.100 to +0.006)
EfficientNet-B3, seed 44	Messidor-2	QWK, grades 0+1 merged	0.201 (0.109–0.298)	-0.002 (-0.061 to +0.054)
EfficientNet-B3, seed 44	Messidor-2	QWK, both corrections	0.106 (0.017–0.204)	-0.037 (-0.121 to +0.036)
EfficientNet-B3, seed 44	IDRiD	QWK	0.210 (0.152–0.273)	+0.058 (+0.016 to +0.101)
EfficientNet-B3, seed 44	IDRiD	QWK, case mix matched	0.255 (0.180–0.331)	+0.069 (+0.016 to +0.125)
EfficientNet-B3, seed 44	IDRiD	QWK, grades 0+1 merged	0.124 (0.044–0.204)	+0.051 (-0.001 to +0.106)
EfficientNet-B3, seed 44	IDRiD	QWK, both corrections	0.121 (0.027–0.214)	+0.074 (+0.010 to +0.142)
EfficientNet-B3, seed 45	Messidor-2	QWK	0.373 (0.309–0.447)	-0.007 (-0.042 to +0.029)
EfficientNet-B3, seed 45	Messidor-2	QWK, case mix matched	0.286 (0.223–0.357)	-0.007 (-0.050 to +0.038)
EfficientNet-B3, seed 45	Messidor-2	QWK, grades 0+1 merged	0.207 (0.116–0.301)	+0.004 (-0.050 to +0.058)
EfficientNet-B3, seed 45	Messidor-2	QWK, both corrections	0.138 (0.043–0.245)	-0.004 (-0.066 to +0.057)
EfficientNet-B3, seed 45	IDRiD	QWK	0.176 (0.125–0.230)	+0.023 (-0.014 to +0.061)
EfficientNet-B3, seed 45	IDRiD	QWK, case mix matched	0.192 (0.136–0.252)	+0.005 (-0.035 to +0.047)
EfficientNet-B3, seed 45	IDRiD	QWK, grades 0+1 merged	0.112 (0.034–0.188)	+0.040 (-0.015 to +0.096)
EfficientNet-B3, seed 45	IDRiD	QWK, both corrections	0.076 (-0.001–0.153)	+0.029 (-0.028 to +0.085)
RETFound LP, seed 43	Messidor-2	QWK	0.467 (0.397–0.543)	+0.086 (+0.023 to +0.148)
RETFound LP, seed 43	Messidor-2	QWK, case mix matched	0.331 (0.268–0.397)	+0.038 (-0.040 to +0.108)
RETFound LP, seed 43	Messidor-2	QWK, grades 0+1 merged	0.293 (0.200–0.390)	+0.091 (+0.008 to +0.172)
RETFound LP, seed 43	Messidor-2	QWK, both corrections	0.147 (0.062–0.237)	+0.004 (-0.093 to +0.094)
RETFound LP, seed 43	IDRiD	QWK	0.282 (0.221–0.340)	+0.129 (+0.071 to +0.187)
RETFound LP, seed 43	IDRiD	QWK, case mix matched	0.381 (0.315–0.440)	+0.195 (+0.123 to +0.260)
RETFound LP, seed 43	IDRiD	QWK, grades 0+1 merged	0.149 (0.070–0.222)	+0.076 (+0.009 to +0.144)
RETFound LP, seed 43	IDRiD	QWK, both corrections	0.214 (0.129–0.292)	+0.166 (+0.087 to +0.242)
RETFound LP, seed 44	Messidor-2	QWK	0.451 (0.381–0.524)	+0.071 (+0.005 to +0.131)
RETFound LP, seed 44	Messidor-2	QWK, case mix matched	0.304 (0.250–0.360)	+0.012 (-0.061 to +0.078)
RETFound LP, seed 44	Messidor-2	QWK, grades 0+1 merged	0.313 (0.218–0.404)	+0.110 (+0.020 to +0.195)
RETFound LP, seed 44	Messidor-2	QWK, both corrections	0.146 (0.071–0.219)	+0.004 (-0.095 to +0.096)
RETFound LP, seed 44	IDRiD	QWK	0.267 (0.208–0.324)	+0.114 (+0.053 to +0.172)
RETFound LP, seed 44	IDRiD	QWK, case mix matched	0.370 (0.304–0.430)	+0.184 (+0.110 to +0.250)
RETFound LP, seed 44	IDRiD	QWK, grades 0+1 merged	0.133 (0.055–0.203)	+0.061 (-0.010 to +0.134)
RETFound LP, seed 44	IDRiD	QWK, both corrections	0.204 (0.119–0.280)	+0.156 (+0.074 to +0.233)
RETFound LP, seed 45	Messidor-2	QWK	0.491 (0.418–0.563)	+0.111 (+0.046 to +0.174)

Model	External	Measure	Drop (95% CI)	Excess drop vs EfficientNet-B3
RETFound LP, seed 45	Messidor-2	QWK, case mix matched	0.342 (0.284–0.401)	+0.049 (-0.027 to +0.121)
RETFound LP, seed 45	Messidor-2	QWK, grades 0+1 merged	0.377 (0.281–0.470)	+0.174 (+0.085 to +0.261)
RETFound LP, seed 45	Messidor-2	QWK, both corrections	0.199 (0.121–0.277)	+0.057 (-0.053 to +0.158)
RETFound LP, seed 45	IDRiD	QWK	0.256 (0.197–0.313)	+0.103 (+0.044 to +0.163)
RETFound LP, seed 45	IDRiD	QWK, case mix matched	0.357 (0.294–0.418)	+0.171 (+0.099 to +0.240)
RETFound LP, seed 45	IDRiD	QWK, grades 0+1 merged	0.123 (0.045–0.194)	+0.050 (-0.022 to +0.123)
RETFound LP, seed 45	IDRiD	QWK, both corrections	0.184 (0.104–0.259)	+0.137 (+0.056 to +0.216)

Referable-DR AUC compared with EfficientNet-B3 (paired bootstrap, Holm adjustment within dataset):

Model	Dataset	Difference (95% CI)	p	Holm p
ResNet-50	APTOS internal	-0.001 (-0.008 to +0.005)	0.676	1.000
ResNet-50	Messidor-2	-0.053 (-0.075 to -0.031)	<0.001	<0.001
ResNet-50	IDRiD	-0.014 (-0.032 to +0.003)	0.113	0.339
RETFound LP	APTOS internal	-0.003 (-0.014 to +0.006)	0.517	1.000
RETFound LP	Messidor-2	-0.060 (-0.090 to -0.032)	<0.001	<0.001
RETFound LP	IDRiD	-0.044 (-0.071 to -0.021)	<0.001	<0.001
RETFound full	APTOS internal	-0.015 (-0.032 to +0.001)	0.069	0.276
RETFound full	Messidor-2	-0.025 (-0.049 to -0.002)	0.033	0.165
RETFound full	IDRiD	-0.051 (-0.081 to -0.019)	0.001	0.006
EfficientNet-B3 + label smoothing (control)	APTOS internal	-0.042 (-0.065 to -0.022)	<0.001	–
EfficientNet-B3 + label smoothing (control)	Messidor-2	-0.048 (-0.068 to -0.029)	<0.001	–
EfficientNet-B3 + label smoothing (control)	IDRiD	-0.035 (-0.053 to -0.018)	<0.001	–
EfficientNet-B3, RETFound recipe (control)	APTOS internal	-0.017 (-0.032 to -0.003)	0.016	–
EfficientNet-B3, RETFound recipe (control)	Messidor-2	-0.032 (-0.049 to -0.015)	<0.001	–
EfficientNet-B3, RETFound recipe (control)	IDRiD	-0.027 (-0.045 to -0.010)	0.004	–
RETFound full, seed 43	APTOS internal	-0.001 (-0.012 to +0.008)	0.946	–
RETFound full, seed 43	Messidor-2	-0.013 (-0.037 to +0.010)	0.240	–
RETFound full, seed 43	IDRiD	-0.005 (-0.028 to +0.019)	0.696	–
RETFound full, seed 44	APTOS internal	-0.008 (-0.023 to +0.006)	0.284	–
RETFound full, seed 44	Messidor-2	-0.043 (-0.072 to -0.016)	0.002	–
RETFound full, seed 44	IDRiD	-0.023 (-0.050 to +0.002)	0.072	–
RETFound full, CNN recipe (control)	APTOS internal	0.000 (-0.007 to +0.007)	0.961	–
RETFound full, CNN recipe (control)	Messidor-2	-0.127 (-0.156 to -0.099)	<0.001	–
RETFound full, CNN recipe (control)	IDRiD	+0.005 (-0.013 to +0.023)	0.593	–
EfficientNet-B3, seed 43	APTOS internal	+0.007 (+0.003 to +0.012)	0.001	–
EfficientNet-B3, seed 43	Messidor-2	+0.001 (-0.011 to +0.013)	0.866	–
EfficientNet-B3, seed 43	IDRiD	+0.006 (-0.007 to +0.018)	0.363	–
EfficientNet-B3, seed 44	APTOS internal	-0.003 (-0.008 to +0.002)	0.281	–
EfficientNet-B3, seed 44	Messidor-2	+0.005 (-0.007 to +0.017)	0.426	–
EfficientNet-B3, seed 44	IDRiD	-0.006 (-0.019 to +0.006)	0.299	–
EfficientNet-B3, seed 45	APTOS internal	+0.001 (-0.004 to +0.006)	0.655	–
EfficientNet-B3, seed 45	Messidor-2	+0.013 (+0.001 to +0.026)	0.032	–
EfficientNet-B3, seed 45	IDRiD	0.000 (-0.011 to +0.012)	0.956	–
RETFound LP, seed 43	APTOS internal	-0.003 (-0.013 to +0.006)	0.560	–
RETFound LP, seed 43	Messidor-2	-0.062 (-0.092 to -0.032)	<0.001	–
RETFound LP, seed 43	IDRiD	-0.062 (-0.092 to -0.034)	<0.001	–
RETFound LP, seed 44	APTOS internal	-0.006 (-0.018 to +0.004)	0.289	–
RETFound LP, seed 44	Messidor-2	-0.053 (-0.082 to -0.025)	<0.001	–
RETFound LP, seed 44	IDRiD	-0.054 (-0.082 to -0.028)	<0.001	–
RETFound LP, seed 45	APTOS internal	-0.004 (-0.016 to +0.006)	0.439	–
RETFound LP, seed 45	Messidor-2	-0.064 (-0.094 to -0.035)	<0.001	–
RETFound LP, seed 45	IDRiD	-0.055 (-0.083 to -0.029)	<0.001	–

S9 Uncertainty, error detection and referral

Error-detection AUC (incorrect versus correct grade) for maximum softmax probability (MSP), top-two margin and predictive entropy, and the difference from MSP.

Model	Dataset	MSP	Margin	Entropy	Margin − MSP	Entropy − MSP
EfficientNet-B3	APTOS internal	0.850 (0.818–0.881)	0.849 (0.817–0.880)	0.850 (0.819–0.881)	-0.001 (-0.004 to +0.002)	0.000 (-0.005 to +0.005)
EfficientNet-B3	Messidor-2	0.682 (0.655–0.710)	0.682 (0.655–0.710)	0.681 (0.654–0.709)	0.000 (-0.001 to +0.001)	0.000 (-0.002 to +0.001)
EfficientNet-B3	IDRiD	0.585 (0.536–0.636)	0.578 (0.531–0.629)	0.585 (0.538–0.638)	-0.006 (-0.012 to 0.000)	+0.001 (-0.013 to +0.014)
ResNet-50	APTOS internal	0.804 (0.769–0.836)	0.810 (0.776–0.843)	0.793 (0.757–0.826)	+0.006 (+0.001 to +0.012)	-0.011 (-0.020 to -0.001)
ResNet-50	Messidor-2	0.654 (0.626–0.681)	0.655 (0.627–0.683)	0.652 (0.625–0.680)	+0.001 (0.000 to +0.003)	-0.001 (-0.003 to +0.001)
ResNet-50	IDRiD	0.499 (0.449–0.551)	0.500 (0.449–0.552)	0.500 (0.451–0.551)	+0.001 (-0.009 to +0.010)	0.000 (-0.016 to +0.018)
RETFound LP	APTOS internal	0.779 (0.734–0.821)	0.790 (0.747–0.830)	0.711 (0.663–0.760)	+0.011 (0.000 to +0.022)	-0.068 (-0.091 to -0.045)
RETFound LP	Messidor-2	0.655 (0.624–0.685)	0.651 (0.620–0.680)	0.649 (0.617–0.680)	-0.004 (-0.012 to +0.003)	-0.006 (-0.015 to +0.003)
RETFound LP	IDRiD	0.579 (0.527–0.632)	0.541 (0.489–0.597)	0.600 (0.548–0.650)	-0.038 (-0.064 to -0.012)	+0.021 (-0.005 to +0.045)
RETFound full	APTOS internal	0.640 (0.577–0.702)	0.666 (0.605–0.726)	0.516 (0.454–0.580)	+0.027 (+0.011 to +0.046)	-0.123 (-0.160 to -0.089)
RETFound full	Messidor-2	0.653 (0.620–0.682)	0.659 (0.628–0.688)	0.633 (0.600–0.663)	+0.006 (0.000 to +0.013)	-0.020 (-0.030 to -0.011)
RETFound full	IDRiD	0.518 (0.469–0.568)	0.525 (0.473–0.575)	0.503 (0.454–0.555)	+0.006 (-0.010 to +0.023)	-0.015 (-0.038 to +0.007)
EfficientNet-B3 + label smoothing (control)	APTOS internal	0.659 (0.596–0.720)	0.683 (0.622–0.744)	0.505 (0.443–0.563)	+0.024 (+0.011 to +0.039)	-0.154 (-0.191 to -0.118)
EfficientNet-B3 + label smoothing (control)	Messidor-2	0.639 (0.611–0.666)	0.629 (0.602–0.657)	0.639 (0.611–0.666)	-0.009 (-0.014 to -0.004)	0.000 (-0.004 to +0.005)
EfficientNet-B3 + label smoothing (control)	IDRiD	0.467 (0.419–0.520)	0.469 (0.420–0.523)	0.452 (0.404–0.504)	+0.002 (-0.012 to +0.016)	-0.016 (-0.040 to +0.007)
EfficientNet-B3, RETFound recipe (control)	APTOS internal	0.707 (0.650–0.765)	0.740 (0.682–0.796)	0.584 (0.524–0.645)	+0.032 (+0.016 to +0.049)	-0.124 (-0.157 to -0.093)
EfficientNet-B3, RETFound recipe (control)	Messidor-2	0.660 (0.632–0.688)	0.660 (0.630–0.688)	0.658 (0.629–0.686)	-0.001 (-0.006 to +0.004)	-0.002 (-0.007 to +0.003)
EfficientNet-B3, RETFound recipe (control)	IDRiD	0.530 (0.479–0.581)	0.523 (0.473–0.573)	0.535 (0.486–0.586)	-0.008 (-0.021 to +0.005)	+0.004 (-0.018 to +0.026)
RETFound full, CNN recipe (control)	APTOS internal	0.847 (0.818–0.875)	0.844 (0.815–0.871)	0.841 (0.812–0.869)	-0.003 (-0.008 to +0.002)	-0.006 (-0.016 to +0.003)
RETFound full, CNN recipe (control)	Messidor-2	0.591 (0.559–0.621)	0.588 (0.557–0.619)	0.598 (0.566–0.628)	-0.003 (-0.004 to -0.001)	+0.007 (+0.004 to +0.010)
RETFound full, CNN recipe (control)	IDRiD	0.610 (0.562–0.659)	0.599 (0.552–0.649)	0.621 (0.574–0.670)	-0.011 (-0.021 to 0.000)	+0.012 (-0.009 to +0.031)
EfficientNet-B3, seed 43	APTOS internal	0.833 (0.802–0.864)	0.837 (0.806–0.866)	0.827 (0.795–0.855)	+0.004 (0.000 to +0.007)	-0.007 (-0.014 to +0.001)
EfficientNet-B3, seed 43	Messidor-2	0.685 (0.657–0.714)	0.685 (0.657–0.714)	0.685 (0.657–0.714)	0.000 (-0.001 to +0.001)	0.000 (-0.001 to +0.001)
EfficientNet-B3, seed 43	IDRiD	0.639 (0.589–0.686)	0.642 (0.594–0.690)	0.637 (0.591–0.684)	+0.003 (-0.004 to +0.011)	-0.002 (-0.015 to +0.010)
RETFound LP, seed 43	APTOS internal	0.767 (0.720–0.808)	0.776 (0.731–0.816)	0.712 (0.661–0.758)	+0.009 (-0.003 to +0.022)	-0.055 (-0.077 to -0.033)
RETFound LP, seed 43	Messidor-2	0.676 (0.647–0.704)	0.664 (0.635–0.692)	0.660 (0.630–0.687)	-0.012 (-0.022 to -0.001)	-0.016 (-0.029 to -0.005)
RETFound LP, seed 43	IDRiD	0.540 (0.487–0.593)	0.503 (0.451–0.555)	0.576 (0.519–0.631)	-0.037 (-0.063 to -0.011)	+0.036 (+0.012 to +0.061)
RETFound full, seed 43	APTOS internal	0.727 (0.670–0.784)	0.743 (0.688–0.798)	0.627 (0.567–0.688)	+0.016 (+0.002 to +0.032)	-0.101 (-0.132 to -0.070)
RETFound full, seed 43	Messidor-2	0.673 (0.643–0.703)	0.672 (0.642–0.701)	0.664 (0.633–0.693)	-0.002 (-0.006 to +0.003)	-0.009 (-0.017 to -0.002)
RETFound full, seed 43	IDRiD	0.639 (0.589–0.684)	0.634 (0.585–0.680)	0.603 (0.552–0.649)	-0.005 (-0.029 to +0.020)	-0.036 (-0.062 to -0.011)
EfficientNet-B3, seed 44	APTOS internal	0.821 (0.780–0.858)	0.822 (0.782–0.858)	0.818 (0.778–0.854)	+0.001 (-0.002 to +0.003)	-0.003 (-0.008 to +0.001)
EfficientNet-B3, seed 44	Messidor-2	0.687 (0.659–0.715)	0.685 (0.657–0.713)	0.688 (0.660–0.716)	-0.002 (-0.004 to 0.000)	0.000 (-0.002 to +0.003)
EfficientNet-B3, seed 44	IDRiD	0.514 (0.465–0.568)	0.511 (0.462–0.565)	0.516 (0.468–0.569)	-0.002 (-0.008 to +0.003)	+0.002 (-0.008 to +0.013)
RETFound LP, seed 44	APTOS internal	0.746 (0.700–0.791)	0.750 (0.705–0.794)	0.678 (0.628–0.728)	+0.004 (-0.010 to +0.017)	-0.068 (-0.090 to -0.045)
RETFound LP, seed 44	Messidor-2	0.681 (0.653–0.710)	0.674 (0.645–0.702)	0.673 (0.643–0.702)	-0.008 (-0.017 to +0.001)	-0.008 (-0.018 to +0.002)
RETFound LP, seed 44	IDRiD	0.556 (0.502–0.608)	0.527 (0.471–0.579)	0.584 (0.528–0.637)	-0.030 (-0.055 to -0.005)	+0.028 (+0.002 to +0.053)
RETFound full, seed 44	APTOS internal	0.700 (0.640–0.756)	0.723 (0.666–0.774)	0.607 (0.544–0.666)	+0.023 (+0.007 to +0.041)	-0.093 (-0.127 to -0.063)
RETFound full, seed 44	Messidor-2	0.652 (0.621–0.682)	0.649 (0.618–0.679)	0.650 (0.618–0.679)	-0.003 (-0.008 to +0.002)	-0.003 (-0.009 to +0.004)
RETFound full, seed 44	IDRiD	0.532 (0.481–0.583)	0.506 (0.452–0.558)	0.549 (0.500–0.601)	-0.026 (-0.048 to -0.006)	+0.017 (-0.008 to +0.043)
EfficientNet-B3, seed 45	APTOS internal	0.850 (0.817–0.880)	0.851 (0.818–0.881)	0.848 (0.815–0.877)	+0.001 (-0.001 to +0.003)	-0.002 (-0.007 to +0.002)
EfficientNet-B3, seed 45	Messidor-2	0.708 (0.680–0.736)	0.708 (0.680–0.736)	0.707 (0.680–0.735)	0.000 (-0.001 to +0.002)	0.000 (-0.002 to +0.001)
EfficientNet-B3, seed 45	IDRiD	0.573 (0.524–0.625)	0.574 (0.525–0.626)	0.563 (0.515–0.617)	+0.002 (-0.004 to +0.007)	-0.009 (-0.021 to +0.002)
RETFound LP, seed 45	APTOS internal	0.770 (0.723–0.815)	0.780 (0.735–0.824)	0.691 (0.640–0.741)	+0.010 (-0.002 to +0.022)	-0.079 (-0.104 to -0.056)
RETFound LP, seed 45	Messidor-2	0.660 (0.630–0.690)	0.650 (0.620–0.679)	0.664 (0.633–0.694)	-0.010 (-0.019 to -0.001)	+0.004 (-0.006 to +0.015)
RETFound LP, seed 45	IDRiD	0.537 (0.483–0.591)	0.488 (0.435–0.541)	0.581 (0.526–0.633)	-0.049 (-0.076 to -0.023)	+0.044 (+0.018 to +0.070)

Referral of the least confident images to a human grader: fraction of all referable-DR images missed by the system (retained and graded non-referable), %, and accuracy and QWK on retained images (MSP ranking). Random: fixed random referral order.

Model	Dataset	Missed, no referral	MSP 10%	MSP 20%	MSP 30%	Entropy 30%	Random 30%	Accuracy retained (30%)	QWK retained (30%)
EfficientNet-B3	Messidor-2	61.3 (56.0–66.7)	50.1 (45.3–54.9)	38.9 (34.6–43.2)	28.2 (24.4–32.6)	28.4 (24.3–32.6)	42.0 (37.3–46.6)	71.4 (68.1–74.4)	0.188 (0.059–0.299)
EfficientNet-B3	IDRiD	17.3 (13.4–21.5)	16.1 (12.3–20.2)	14.9 (11.4–18.8)	12.7 (8.9–16.2)	13.0 (9.3–16.4)	13.3 (9.6–17.0)	50.1 (45.2–55.4)	0.742 (0.698–0.785)
EfficientNet-B3	APTOS internal	10.4 (7.0–14.3)	7.5 (4.3–10.5)	3.4 (1.4–5.7)	1.5 (0.0–3.3)	1.9 (0.4–3.8)	7.5 (4.5–10.9)	92.4 (89.7–94.8)	0.949 (0.926–0.969)

Model	Dataset	Missed, no referral	MSP 10%	MSP 20%	MSP 30%	Entropy 30%	Random 30%	Accuracy retained (30%)	QWK retained (30%)
ResNet-50	Messidor-2	64.3 (58.7–69.6)	55.8 (50.5–60.6)	47.0 (42.7–51.4)	35.7 (31.2–39.5)	35.0 (30.9–39.0)	44.9 (39.7–49.8)	68.1 (64.9–71.4)	0.072 (0.000–0.146)
ResNet-50	IDRiD	28.5 (23.5–33.2)	25.4 (20.5–29.9)	22.9 (18.6–27.5)	20.7 (16.6–25.1)	22.3 (17.7–26.4)	20.7 (16.4–25.0)	42.1 (37.1–47.4)	0.701 (0.643–0.754)
ResNet-50	APTOS internal	13.4 (9.6–18.1)	11.6 (8.0–15.7)	9.0 (5.9–12.5)	6.3 (3.6–9.8)	7.8 (4.6–11.1)	9.3 (6.1–13.1)	89.7 (86.0–92.1)	0.913 (0.871–0.942)
RETFound LP	Messidor-2	57.1 (51.3–62.9)	47.3 (41.9–52.9)	39.2 (34.3–44.3)	33.5 (28.7–38.1)	30.0 (25.4–35.0)	40.5 (35.3–45.4)	64.8 (61.6–68.1)	0.450 (0.353–0.539)
RETFound LP	IDRiD	17.3 (13.3–21.4)	15.2 (11.4–19.0)	13.9 (10.5–17.7)	11.8 (8.3–15.3)	10.8 (7.8–14.4)	13.9 (10.2–17.5)	33.0 (28.5–38.8)	0.700 (0.649–0.747)
RETFound LP	APTOS internal	15.3 (11.4–19.7)	10.8 (7.4–14.5)	6.7 (4.1–10.7)	4.5 (2.2–7.0)	4.1 (1.8–6.2)	10.8 (7.4–14.8)	88.2 (85.3–91.3)	0.924 (0.895–0.954)
RETFound full	Messidor-2	60.8 (55.3–66.0)	46.8 (41.8–52.2)	37.4 (33.0–42.2)	26.5 (22.3–30.9)	25.4 (20.9–29.9)	43.1 (38.0–47.4)	69.9 (66.5–72.9)	0.556 (0.466–0.627)
RETFound full	IDRiD	26.6 (21.7–31.5)	23.2 (18.5–27.8)	20.7 (16.0–24.7)	18.9 (14.7–23.1)	18.3 (13.9–22.7)	21.1 (16.5–25.1)	41.6 (36.6–46.8)	0.753 (0.706–0.796)
RETFound full	APTOS internal	10.1 (6.9–13.9)	7.5 (4.7–10.8)	4.5 (2.3–7.4)	3.4 (1.4–5.8)	5.2 (2.7–8.0)	6.3 (3.6–9.6)	87.8 (84.9–90.7)	0.931 (0.906–0.954)
EfficientNet-B3 + label smoothing (control)	Messidor-2	65.6 (60.2–70.9)	54.0 (49.0–58.9)	41.6 (37.6–46.7)	32.8 (28.5–37.1)	31.5 (27.4–36.0)	45.3 (40.6–50.1)	69.0 (66.0–72.2)	0.292 (0.161–0.406)
EfficientNet-B3 + label smoothing (control)	IDRiD	26.0 (21.2–30.8)	23.8 (19.2–28.6)	22.6 (18.0–27.2)	19.2 (15.2–23.7)	18.9 (14.5–23.0)	20.4 (16.3–24.8)	42.4 (36.8–47.4)	0.765 (0.704–0.796)
EfficientNet-B3 + label smoothing (control)	APTOS internal	16.8 (12.5–21.3)	11.9 (8.3–16.2)	10.1 (6.5–13.7)	8.2 (5.1–11.6)	11.2 (7.8–15.5)	12.3 (8.7–16.5)	87.4 (84.5–90.5)	0.912 (0.879–0.941)
EfficientNet-B3, RETFound recipe (control)	Messidor-2	62.8 (57.5–68.1)	50.5 (45.9–55.7)	38.7 (34.6–43.1)	28.0 (24.3–32.6)	28.2 (23.9–32.4)	42.9 (38.2–47.5)	70.0 (66.6–72.9)	0.286 (0.164–0.385)
EfficientNet-B3, RETFound recipe (control)	IDRiD	24.5 (19.8–29.0)	21.7 (17.3–26.5)	18.6 (14.4–22.5)	15.5 (12.1–19.6)	16.4 (12.3–20.1)	18.6 (14.6–23.0)	44.0 (39.1–49.6)	0.717 (0.667–0.765)
EfficientNet-B3, RETFound recipe (control)	APTOS internal	13.8 (10.0–18.2)	10.8 (7.3–14.5)	7.5 (4.6–10.8)	5.2 (2.8–8.3)	9.0 (5.9–12.7)	10.4 (7.0–14.3)	89.5 (86.4–92.1)	0.916 (0.878–0.945)
RETFound full, CNN recipe (control)	Messidor-2	56.2 (51.0–61.7)	52.7 (47.6–58.0)	49.0 (43.8–54.3)	45.7 (40.8–50.6)	46.2 (41.3–51.0)	38.5 (33.8–43.0)	63.1 (59.8–66.4)	0.235 (0.122–0.339)
RETFound full, CNN recipe (control)	IDRiD	21.1 (16.6–25.7)	18.9 (14.9–23.6)	17.6 (13.6–21.9)	16.4 (12.3–20.2)	14.6 (10.9–18.3)	15.8 (11.8–19.7)	53.5 (49.0–59.0)	0.783 (0.740–0.829)
RETFound full, CNN recipe (control)	APTOS internal	10.8 (7.4–14.9)	8.2 (5.3–11.9)	6.7 (3.9–9.7)	4.5 (2.3–7.3)	5.2 (2.9–8.1)	6.7 (4.1–10.1)	89.9 (86.4–92.6)	0.934 (0.899–0.957)
EfficientNet-B3, seed 43	Messidor-2	65.0 (59.3–70.4)	56.0 (50.8–61.1)	43.5 (39.1–47.3)	29.5 (25.7–34.1)	29.8 (25.8–34.0)	46.0 (40.8–50.9)	70.0 (66.6–73.1)	0.000 (0.000–0.000)
EfficientNet-B3, seed 43	IDRiD	22.3 (17.8–27.1)	20.4 (16.1–25.0)	18.0 (13.8–22.1)	15.5 (11.3–19.2)	16.1 (12.5–20.2)	17.3 (13.4–21.5)	51.5 (46.8–57.3)	0.787 (0.748–0.823)
EfficientNet-B3, seed 43	APTOS internal	13.8 (9.8–18.2)	10.1 (6.7–13.9)	7.5 (4.5–10.6)	3.4 (1.2–5.6)	3.4 (1.4–6.1)	9.7 (6.4–13.7)	91.1 (88.6–93.8)	0.923 (0.889–0.952)
RETFound LP, seed 43	Messidor-2	60.2 (54.6–65.7)	51.2 (45.9–56.5)	44.4 (39.4–49.6)	35.7 (31.2–40.8)	35.2 (30.4–40.0)	42.5 (37.4–46.9)	62.6 (58.9–65.9)	0.464 (0.358–0.548)
RETFound LP, seed 43	IDRiD	21.1 (16.6–25.6)	18.3 (14.2–22.4)	16.1 (12.4–20.6)	13.0 (9.6–16.7)	12.7 (8.9–16.1)	16.4 (12.3–20.3)	33.8 (28.5–38.5)	0.682 (0.626–0.725)
RETFound LP, seed 43	APTOS internal	13.1 (9.5–17.2)	10.4 (7.2–14.3)	8.2 (4.9–11.5)	4.5 (2.1–7.0)	4.5 (1.9–6.8)	9.0 (6.0–12.7)	88.8 (85.7–91.7)	0.923 (0.887–0.948)
RETFound full, seed 43	Messidor-2	68.7 (63.2–73.8)	56.7 (51.8–61.6)	44.9 (40.6–49.8)	33.0 (28.9–37.5)	30.2 (26.2–35.1)	49.5 (44.1–53.8)	69.4 (65.9–72.2)	0.358 (0.234–0.466)
RETFound full, seed 43	IDRiD	16.4 (12.4–20.6)	13.6 (10.1–17.6)	11.8 (8.2–15.3)	9.3 (5.8–12.4)	8.4 (5.5–11.7)	12.7 (9.1–16.3)	52.9 (47.4–58.2)	0.786 (0.734–0.830)
RETFound full, seed 43	APTOS internal	7.1 (4.2–10.4)	3.4 (1.5–5.9)	2.2 (0.7–4.2)	1.5 (0.4–3.4)	2.6 (0.7–4.5)	4.9 (2.6–7.7)	90.9 (88.0–93.6)	0.946 (0.921–0.966)
EfficientNet-B3, seed 44	Messidor-2	55.6 (50.1–60.8)	44.2 (39.4–49.0)	34.1 (29.8–38.5)	27.6 (23.5–31.6)	27.6 (23.5–31.6)	37.9 (33.2–42.6)	70.4 (67.1–73.5)	0.360 (0.261–0.467)
EfficientNet-B3, seed 44	IDRiD	19.2 (15.0–23.4)	17.6 (13.4–21.6)	17.0 (12.9–20.8)	14.2 (10.8–18.0)	15.2 (11.3–18.6)	13.0 (9.6–16.8)	41.6 (36.8–46.8)	0.711 (0.656–0.760)
EfficientNet-B3, seed 44	APTOS internal	10.1 (6.8–13.8)	6.0 (3.3–9.0)	4.5 (2.0–6.9)	3.0 (1.2–5.3)	3.0 (1.2–5.3)	7.5 (4.4–10.6)	91.9 (89.3–94.6)	0.923 (0.891–0.955)
RETFound LP, seed 44	Messidor-2	56.2 (50.4–61.9)	46.4 (40.8–51.4)	38.9 (33.8–43.8)	30.4 (25.9–35.1)	29.3 (25.2–34.3)	40.3 (35.2–44.7)	64.1 (60.6–67.5)	0.475 (0.380–0.560)
RETFound LP, seed 44	IDRiD	20.1 (15.7–24.6)	17.3 (13.4–21.5)	15.5 (11.8–19.8)	13.6 (10.0–17.3)	13.0 (9.3–16.9)	16.1 (12.2–19.9)	32.4 (27.7–37.4)	0.681 (0.628–0.732)
RETFound LP, seed 44	APTOS internal	19.8 (15.2–24.6)	14.2 (10.6–18.6)	10.8 (7.5–14.6)	6.0 (3.5–9.5)	7.1 (4.1–10.5)	13.8 (10.1–18.3)	87.6 (84.1–90.5)	0.925 (0.887–0.949)
RETFound full, seed 44	Messidor-2	72.4 (67.3–77.5)	57.8 (52.8–62.6)	46.4 (41.8–51.5)	36.8 (32.4–41.2)	35.0 (30.9–39.9)	51.9 (46.5–56.0)	68.1 (64.9–71.2)	0.319 (0.201–0.425)
RETFound full, seed 44	IDRiD	34.1 (28.7–39.2)	29.4 (24.6–34.7)	24.5 (20.1–29.3)	21.7 (17.2–25.9)	18.6 (14.6–22.7)	26.0 (21.0–30.5)	43.2 (38.2–48.5)	0.781 (0.740–0.817)
RETFound full, seed 44	APTOS internal	11.2 (7.5–15.1)	7.5 (4.3–10.7)	4.9 (2.4–7.6)	3.4 (1.4–5.7)	4.5 (2.1–7.1)	7.1 (4.3–10.3)	90.3 (87.4–92.8)	0.940 (0.912–0.962)
EfficientNet-B3, seed 45	Messidor-2	60.4 (55.0–65.9)	48.8 (43.7–53.4)	40.0 (35.1–44.0)	28.7 (24.6–32.9)	28.9 (24.6–32.8)	41.6 (36.5–46.2)	70.6 (67.1–73.6)	0.333 (0.237–0.427)
EfficientNet-B3, seed 45	IDRiD	22.6 (18.0–27.2)	20.4 (15.8–24.7)	16.7 (12.9–21.1)	14.9 (11.1–18.6)	15.5 (11.8–19.4)	16.4 (12.6–20.5)	43.8 (38.5–48.8)	0.765 (0.724–0.806)
EfficientNet-B3, seed 45	APTOS internal	12.3 (8.6–16.3)	9.7 (6.2–13.1)	5.6 (2.9–8.5)	3.7 (1.6–6.2)	3.7 (1.6–6.2)	8.2 (5.3–11.9)	92.8 (89.9–95.0)	0.936 (0.899–0.964)
RETFound LP, seed 45	Messidor-2	53.0 (47.4–58.5)	45.1 (39.7–50.4)	37.9 (33.0–43.1)	31.3 (26.2–35.6)	28.4 (23.9–32.9)	37.2 (32.6–41.8)	63.6 (60.4–66.9)	0.425 (0.330–0.510)
RETFound LP, seed 45	IDRiD	22.3 (17.6–26.7)	19.2 (14.9–23.4)	17.0 (12.9–20.8)	14.6 (10.8–18.2)	13.6 (10.0–17.6)	18.3 (14.2–22.2)	31.9 (27.1–37.1)	0.675 (0.628–0.730)
RETFound LP, seed 45	APTOS internal	16.8 (12.7–21.5)	12.7 (8.9–17.0)	7.8 (4.4–11.2)	5.2 (2.4–7.8)	4.9 (2.5–8.0)	11.9 (8.4–16.4)	88.4 (85.5–91.5)	0.929 (0.898–0.956)

S10 Saliency localization against expert lesion masks

IDRiD Segmentation subset, 81 images. Lift = share of saliency mass on lesions divided by lesion share of the field of view. Cross-image = the model's map for a different image (derangement, seed 42).

Lift MA, HE, EX and SE: lift computed separately for microaneurysms, hemorrhages, hard exudates and soft exudates. Energy on OD: share of saliency mass on the optic disc. Values in brackets are 95% bootstrap intervals.

Map	Energy on lesions	Lift (all lesions)	Pointing hit	Pixel AUC	Top-10% precision	Lift MA	Lift HE	Lift EX	Lift SE	Energy on OD
Center-Gaussian control	0.09 [0.07, 0.12]	1.28 [1.19, 1.38]	0.36 [0.25, 0.47]	0.59 [0.56, 0.61]	0.11 [0.08, 0.14]	1.23	0.90	1.80	0.70	0.03
EfficientNet-B3 (cross-image)	0.08 [0.06, 0.11]	1.17 [1.08, 1.25]	0.31 [0.21, 0.41]	0.55 [0.53, 0.58]	0.09 [0.06, 0.12]	1.18	0.91	1.49	0.70	0.03
ResNet-50 (cross-image)	0.08 [0.06, 0.11]	1.15 [1.08, 1.22]	0.33 [0.23, 0.44]	0.56 [0.53, 0.58]	0.09 [0.07, 0.12]	1.18	0.88	1.46	0.80	0.03
RETFound full fine-tune (cross-image)	0.07 [0.05, 0.10]	1.02 [0.94, 1.11]	0.25 [0.16, 0.35]	0.51 [0.49, 0.53]	0.08 [0.05, 0.11]	1.08	0.91	1.13	0.81	0.04
RETFound linear probe (cross-image)	0.08 [0.06, 0.11]	1.12 [1.04, 1.21]	0.36 [0.26, 0.47]	0.55 [0.52, 0.57]	0.09 [0.06, 0.12]	1.10	1.03	1.30	0.83	0.03
EfficientNet-B3	0.09 [0.07, 0.13]	1.37 [1.28, 1.48]	0.48 [0.37, 0.59]	0.61 [0.58, 0.63]	0.12 [0.09, 0.15]	1.21	1.03	1.73	0.87	0.03
ResNet-50	0.09 [0.07, 0.12]	1.32 [1.25, 1.41]	0.49 [0.38, 0.60]	0.61 [0.59, 0.64]	0.12 [0.09, 0.15]	1.21	1.02	1.61	0.95	0.03
RETFound full fine-tune	0.12 [0.09, 0.15]	1.79 [1.63, 1.95]	0.64 [0.53, 0.74]	0.67 [0.63, 0.70]	0.19 [0.15, 0.23]	1.24	1.46	1.83	2.24	0.02
RETFound linear probe	0.12 [0.09, 0.16]	1.67 [1.54, 1.82]	0.60 [0.51, 0.72]	0.67 [0.64, 0.70]	0.19 [0.15, 0.23]	1.20	1.46	1.74	1.77	0.02

Mean lesion share of the field of view: 0.071.

Model	vs	metric	model	comparator	Δ	images better	p

Paired comparisons (Wilcoxon signed-rank over images)

Model	vs	metric	model	comparator	Δ	images better	p
EfficientNet-B3	center	lift	1.372	1.283	+0.090	0.57	0.099
EfficientNet-B3	center	Pixel AUC	0.606	0.587	+0.020	0.55	0.082
EfficientNet-B3	center	energy (lesions)	0.094	0.090	+0.004	0.57	0.068
EfficientNet-B3	center	Top-10% precision	0.121	0.107	+0.013	0.53	0.18
EfficientNet-B3	center	energy (optic disc)	0.026	0.027	-0.001	0.42	0.62
EfficientNet-B3	cross-image	lift	1.372	1.165	+0.207	0.77	1.6e-08
EfficientNet-B3	cross-image	Pixel AUC	0.606	0.555	+0.051	0.75	2.3e-08
EfficientNet-B3	cross-image	energy (lesions)	0.094	0.082	+0.012	0.77	1.2e-08
EfficientNet-B3	cross-image	Top-10% precision	0.121	0.091	+0.030	0.72	1.2e-06
EfficientNet-B3	cross-image	energy (optic disc)	0.026	0.032	-0.006	0.38	0.0088
ResNet-50	center	lift	1.324	1.283	+0.041	0.54	0.45
ResNet-50	center	Pixel AUC	0.612	0.587	+0.025	0.62	0.0078
ResNet-50	center	energy (lesions)	0.090	0.090	-0.001	0.54	0.75
ResNet-50	center	Top-10% precision	0.119	0.107	+0.012	0.60	0.024
ResNet-50	center	energy (optic disc)	0.032	0.027	+0.004	0.57	0.023
ResNet-50	cross-image	lift	1.324	1.147	+0.177	0.72	2.1e-07
ResNet-50	cross-image	Pixel AUC	0.612	0.555	+0.057	0.82	4.8e-10
ResNet-50	cross-image	energy (lesions)	0.090	0.080	+0.010	0.72	6.9e-07
ResNet-50	cross-image	Top-10% precision	0.119	0.091	+0.028	0.56	0.0033
ResNet-50	cross-image	energy (optic disc)	0.032	0.031	+0.001	0.51	0.6
ResNet-50	EfficientNet-B3	lift	1.324	1.372	-0.049	0.42	0.17
ResNet-50	EfficientNet-B3	Pixel AUC	0.612	0.606	+0.006	0.51	0.23
ResNet-50	EfficientNet-B3	energy (lesions)	0.090	0.094	-0.004	0.42	0.048
ResNet-50	EfficientNet-B3	Top-10% precision	0.119	0.121	-0.002	0.53	0.8
ResNet-50	EfficientNet-B3	energy (optic disc)	0.032	0.026	+0.005	0.77	1.4e-05
RETFound linear probe	center	lift	1.672	1.283	+0.389	0.73	3e-06
RETFound linear probe	center	Pixel AUC	0.670	0.587	+0.083	0.72	1.1e-05
RETFound linear probe	center	energy (lesions)	0.121	0.090	+0.031	0.73	1.1e-07
RETFound linear probe	center	Top-10% precision	0.187	0.107	+0.079	0.72	8.9e-08
RETFound linear probe	center	energy (optic disc)	0.016	0.027	-0.011	0.30	3.7e-06
RETFound linear probe	cross-image	lift	1.672	1.123	+0.548	0.79	2.9e-09
RETFound linear probe	cross-image	Pixel AUC	0.670	0.547	+0.123	0.80	2.2e-08
RETFound linear probe	cross-image	energy (lesions)	0.121	0.079	+0.042	0.79	3.8e-09
RETFound linear probe	cross-image	Top-10% precision	0.187	0.089	+0.098	0.80	2.7e-09
RETFound linear probe	cross-image	energy (optic disc)	0.016	0.031	-0.015	0.30	1.8e-06
RETFound linear probe	EfficientNet-B3	lift	1.672	1.372	+0.299	0.72	0.00015
RETFound linear probe	EfficientNet-B3	Pixel AUC	0.670	0.606	+0.063	0.71	0.00022
RETFound linear probe	EfficientNet-B3	energy (lesions)	0.121	0.094	+0.027	0.72	2.9e-06
RETFound linear probe	EfficientNet-B3	Top-10% precision	0.187	0.121	+0.066	0.73	2.2e-06
RETFound linear probe	EfficientNet-B3	energy (optic disc)	0.016	0.026	-0.010	0.33	4.9e-05
RETFound full fine-tune	center	lift	1.788	1.283	+0.505	0.69	7e-06
RETFound full fine-tune	center	Pixel AUC	0.668	0.587	+0.082	0.68	0.00014
RETFound full fine-tune	center	energy (lesions)	0.121	0.090	+0.031	0.69	1.1e-05
RETFound full fine-tune	center	Top-10% precision	0.186	0.107	+0.079	0.72	3.3e-06
RETFound full fine-tune	center	energy (optic disc)	0.019	0.027	-0.008	0.36	0.00073
RETFound full fine-tune	cross-image	lift	1.788	1.022	+0.766	0.85	1.1e-10
RETFound full fine-tune	cross-image	Pixel AUC	0.668	0.509	+0.159	0.82	9.1e-10
RETFound full fine-tune	cross-image	energy (lesions)	0.121	0.073	+0.049	0.85	2.5e-11
RETFound full fine-tune	cross-image	Top-10% precision	0.186	0.078	+0.108	0.83	1.7e-11
RETFound full fine-tune	cross-image	energy (optic disc)	0.019	0.036	-0.017	0.26	3.3e-06
RETFound full fine-tune	EfficientNet-B3	lift	1.788	1.372	+0.416	0.65	0.00012
RETFound full fine-tune	EfficientNet-B3	Pixel AUC	0.668	0.606	+0.062	0.66	0.0008
RETFound full fine-tune	EfficientNet-B3	energy (lesions)	0.121	0.094	+0.027	0.65	5.5e-05
RETFound full fine-tune	EfficientNet-B3	Top-10% precision	0.186	0.121	+0.065	0.69	2.8e-05

Model	vs	metric	model	comparator	Δ	images better	p
RETFound full fine-tune	EfficientNet-B3	energy (optic disc)	0.019	0.026	-0.007	0.40	0.0064

Resolution-matched control

All maps were degraded to the CNN's native 7 × 7 grid (average pooling, then bilinear upsampling to 224 × 224).

Model	resolution	Lift (all lesions)	Pixel AUC	Pointing hit	Top-10% precision
EfficientNet-B3	7 × 7	1.32 [1.24, 1.41]	0.61 [0.58, 0.63]	0.47 [0.36, 0.58]	0.12 [0.09, 0.15]
EfficientNet-B3	native	1.37 [1.28, 1.48]	0.61 [0.58, 0.63]	0.48 [0.37, 0.59]	0.12 [0.09, 0.15]
ResNet-50	7 × 7	1.27 [1.20, 1.34]	0.61 [0.58, 0.63]	0.46 [0.36, 0.57]	0.12 [0.09, 0.15]
ResNet-50	native	1.32 [1.25, 1.41]	0.61 [0.59, 0.64]	0.49 [0.38, 0.60]	0.12 [0.09, 0.15]
RETFound full fine-tune	7 × 7	1.44 [1.34, 1.54]	0.66 [0.63, 0.69]	0.54 [0.43, 0.65]	0.16 [0.12, 0.20]
RETFound full fine-tune	native	1.79 [1.63, 1.95]	0.67 [0.63, 0.70]	0.64 [0.53, 0.74]	0.19 [0.15, 0.23]
RETFound linear probe	7 × 7	1.41 [1.33, 1.50]	0.65 [0.62, 0.68]	0.59 [0.49, 0.70]	0.16 [0.13, 0.20]
RETFound linear probe	native	1.67 [1.54, 1.82]	0.67 [0.64, 0.70]	0.60 [0.51, 0.72]	0.19 [0.15, 0.23]

RETFound (7 × 7) vs EfficientNet-B3 (native)	metric	RETFound	EffB3	Δ	images better	p
RETFound full fine-tune	lift	1.440	1.372	+0.067	0.58	0.17
RETFound full fine-tune	Pixel AUC	0.661	0.606	+0.055	0.68	0.00041
RETFound full fine-tune	Top-10% precision	0.157	0.121	+0.036	0.62	0.0029
RETFound linear probe	lift	1.409	1.372	+0.037	0.58	0.13
RETFound linear probe	Pixel AUC	0.650	0.606	+0.044	0.65	0.00095
RETFound linear probe	Top-10% precision	0.160	0.121	+0.039	0.65	0.00021

S11 Earlier sensitivity analyses (run with the third-party IDRiD copy)

These analyses were run before the official IDRiD release was obtained and before the changes in S0; they are reported for completeness. Their IDRiD results are superseded by the main analysis.

S11.1 Main comparison on the third-party IDRiD copy

Confirmatory family (primary): difference in QWK drop vs EfficientNet-B3

Model	External	Δ drop	95% CI	Bonferroni 98.75% CI	p (boot)	p (Holm)
RETFound full fine-tune	Messidor-2	+0.018	[-0.024, 0.058]	[-0.036, 0.068]	0.3938	0.7876
RETFound full fine-tune	IDRiD	+0.022	[-0.033, 0.083]	[-0.051, 0.096]	0.4548	0.7876
RETFound linear probe	Messidor-2	+0.076	[0.019, 0.132]	[0.004, 0.146]	0.0100	0.0300
RETFound linear probe	IDRiD	+0.105	[0.038, 0.170]	[0.021, 0.195]	0.0040	0.0160

Secondary: ResNet-50 vs EfficientNet-B3

External	Δ drop	95% CI	p (boot)
Messidor-2	+0.063	[0.017, 0.109]	0.0050
IDRiD	+0.051	[0.003, 0.102]	0.0410

Paired QWK difference (model - EfficientNet-B3)

Model	Dataset	QWK model	QWK base	diff	95% CI
ResNet-50	APTOS internal	0.875	0.876	-0.001	[-0.021, 0.019]
ResNet-50	Messidor-2	0.435	0.499	-0.064	[-0.107, -0.023]
ResNet-50	IDRiD	0.707	0.760	-0.052	[-0.098, -0.004]
RETFound linear probe	APTOS internal	0.847	0.876	-0.029	[-0.054, -0.005]
RETFound linear probe	Messidor-2	0.394	0.499	-0.105	[-0.157, -0.053]
RETFound linear probe	IDRiD	0.626	0.760	-0.134	[-0.192, -0.075]
RETFound full fine-tune	APTOS internal	0.882	0.876	+0.006	[-0.014, 0.029]
RETFound full fine-tune	Messidor-2	0.487	0.499	-0.011	[-0.047, 0.025]
RETFound full fine-tune	IDRiD	0.744	0.760	-0.016	[-0.068, 0.038]

Worst-case non-independence (m = n/2)

Model	External	Δ drop	95% CI	excludes 0
RETFound full fine-tune	Messidor-2	+0.018	[-0.042, 0.078]	False
RETFound full fine-tune	IDRiD	+0.022	[-0.056, 0.105]	False
RETFound linear probe	Messidor-2	+0.076	[-0.004, 0.157]	False
RETFound linear probe	IDRiD	+0.105	[0.020, 0.199]	True
ResNet-50	Messidor-2	+0.063	[-0.002, 0.126]	False
ResNet-50	IDRiD	+0.051	[-0.020, 0.122]	False

IDRiD reweighted to the official 516-image grade mix

Model	QWK (copy)	QWK reweighted	Δ drop reweighted (95% CI)
EfficientNet-B3	0.760	0.765	—
ResNet-50	0.707	0.722	+0.042 [-0.006, 0.093]
RETFound linear probe	0.626	0.619	+0.117 [0.049, 0.184]
RETFound full fine-tune	0.744	0.750	+0.022 [-0.034, 0.081]

S11.2 Twenty-pass Monte Carlo dropout (stochastic layers active only where active in training)

Confirmatory family (primary): difference in QWK drop vs EfficientNet-B3

Model	External	Δ drop	95% CI	Bonferroni 98.75% CI	p (boot)	p (Holm)
RETFound full fine-tune	Messidor-2	+0.019	[-0.024, 0.062]	[-0.036, 0.073]	0.3988	0.4278
RETFound full fine-tune	IDRiD	+0.035	[-0.020, 0.092]	[-0.035, 0.107]	0.2139	0.4278
RETFound linear probe	Messidor-2	+0.071	[0.015, 0.126]	[0.002, 0.139]	0.0130	0.0390
RETFound linear probe	IDRiD	+0.105	[0.040, 0.169]	[0.020, 0.195]	0.0040	0.0160

Secondary: ResNet-50 vs EfficientNet-B3

External	Δ drop	95% CI	p (boot)
Messidor-2	+0.057	[0.010, 0.108]	0.0210
IDRiD	+0.047	[-0.001, 0.096]	0.0570

Paired QWK difference (model - EfficientNet-B3)

Model	Dataset	QWK model	QWK base	diff	95% CI
ResNet-50	APTOS internal	0.874	0.881	-0.007	[-0.028, 0.015]
ResNet-50	Messidor-2	0.435	0.499	-0.064	[-0.108, -0.022]
ResNet-50	IDRiD	0.710	0.764	-0.054	[-0.099, -0.007]
RETFound linear probe	APTOS internal	0.847	0.881	-0.034	[-0.058, -0.008]
RETFound linear probe	Messidor-2	0.394	0.499	-0.105	[-0.156, -0.054]
RETFound linear probe	IDRiD	0.626	0.764	-0.139	[-0.197, -0.081]
RETFound full fine-tune	APTOS internal	0.887	0.881	+0.006	[-0.013, 0.026]
RETFound full fine-tune	Messidor-2	0.486	0.499	-0.013	[-0.050, 0.027]
RETFound full fine-tune	IDRiD	0.735	0.764	-0.029	[-0.082, 0.022]

Worst-case non-independence (m = n/2)

Model	External	Δ drop	95% CI	excludes 0
RETFound full fine-tune	Messidor-2	+0.019	[-0.040, 0.079]	False
RETFound full fine-tune	IDRiD	+0.035	[-0.043, 0.115]	False
RETFound linear probe	Messidor-2	+0.071	[-0.008, 0.151]	False
RETFound linear probe	IDRiD	+0.105	[0.019, 0.199]	True
ResNet-50	Messidor-2	+0.057	[-0.009, 0.126]	False
ResNet-50	IDRiD	+0.047	[-0.024, 0.117]	False

IDRiD reweighted to the official 516-image grade mix

Model	QWK (copy)	QWK reweighted	Δ drop reweighted (95% CI)
EfficientNet-B3	0.764	0.770	—
ResNet-50	0.710	0.725	+0.039 [-0.008, 0.086]
RETFound linear probe	0.626	0.619	+0.117 [0.051, 0.184]
RETFound full fine-tune	0.735	0.741	+0.035 [-0.022, 0.092]

S11.3 Training seeds

Additional runs scored with the original evaluation code; the primary model is seed 42.

QWK across seeds (mean, SD, minimum, maximum)

Family	Dataset	seeds	QWK mean	SD	min	max	range
EfficientNet-B3	APTOS internal	4	0.879	0.003	0.876	0.883	0.007
EfficientNet-B3	Messidor-2	4	0.500	0.024	0.469	0.527	0.058
EfficientNet-B3	IDRiD	4	0.749	0.015	0.732	0.764	0.033
RETFound LP	APTOS internal	4	0.855	0.008	0.847	0.865	0.018
RETFound LP	Messidor-2	4	0.395	0.015	0.374	0.408	0.034
RETFound LP	IDRiD	4	0.623	0.012	0.605	0.631	0.026

Internal-to-external QWK drop across seeds

Family	External	seeds	drop mean	SD	min	max
EfficientNet-B3	Messidor-2	4	0.379	0.025	0.353	0.413
EfficientNet-B3	IDRiD	4	0.130	0.015	0.116	0.148
RETFound LP	Messidor-2	4	0.461	0.017	0.444	0.484
RETFound LP	IDRiD	4	0.233	0.019	0.221	0.260

S11.4 Loss function and class weighting

QWK by model family, loss and dataset

Family	Loss	Internal	Messidor-2	IDRiD
EfficientNet-B3	weighted CE	0.876	0.499	0.760
EfficientNet-B3	unweighted CE	0.881	0.549	0.706
EfficientNet-B3	focal (γ = 2)	0.877	0.543	0.769
RETFound LP	weighted CE	0.847	0.394	0.626
RETFound LP	unweighted CE	0.851	0.316	0.751
RETFound LP	focal (γ = 2)	0.858	0.404	0.636

Internal-to-external QWK drop by model family and loss

Family	Loss	Messidor-2 drop	IDRiD drop
EfficientNet-B3	weighted CE	0.377	0.116
EfficientNet-B3	unweighted CE	0.332	0.175
EfficientNet-B3	focal (γ = 2)	0.334	0.108
RETFound LP	weighted CE	0.453	0.221
RETFound LP	unweighted CE	0.535	0.100
RETFound LP	focal (γ = 2)	0.454	0.222

S11.5 Inference-time input changes (contrast enhancement, color transfer, resolution)

QWK by model, dataset and input variant:

Model	Dataset	variant	QWK	acc	referable AUC
EfficientNet-B3	APTOS internal	baseline	0.876	0.803	0.980
EfficientNet-B3	Messidor-2	baseline	0.499	0.626	0.842
EfficientNet-B3	IDRiD	baseline	0.760	0.497	0.962
EfficientNet-B3	APTOS internal	clahe	0.807	0.641	0.977
EfficientNet-B3	Messidor-2	clahe	0.458	0.606	0.792
EfficientNet-B3	IDRiD	clahe	0.697	0.448	0.966
EfficientNet-B3	APTOS internal	color transfer	0.816	0.704	0.979
EfficientNet-B3	Messidor-2	color transfer	0.512	0.608	0.809
EfficientNet-B3	IDRiD	color transfer	0.689	0.464	0.960
RETFound full	APTOS internal	baseline	0.882	0.783	0.967
RETFound full	Messidor-2	baseline	0.487	0.624	0.817
RETFound full	IDRiD	baseline	0.744	0.451	0.885
RETFound full	APTOS internal	clahe	0.703	0.570	0.966
RETFound full	Messidor-2	clahe	0.096	0.588	0.733
RETFound full	IDRiD	clahe	0.410	0.380	0.938
RETFound full	APTOS internal	color transfer	0.810	0.689	0.944
RETFound full	Messidor-2	color transfer	0.284	0.608	0.742
RETFound full	IDRiD	color transfer	0.693	0.505	0.936

Internal-to-external QWK drop, and change versus the unmodified input (negative values mean the variant recovers part of the drop):

Model	External	variant	QWK drop	delta vs baseline
EfficientNet-B3	Messidor-2	baseline	0.377	+0.000
EfficientNet-B3	Messidor-2	clahe	0.350	-0.028
EfficientNet-B3	Messidor-2	color transfer	0.304	-0.073
EfficientNet-B3	IDRiD	baseline	0.116	+0.000
EfficientNet-B3	IDRiD	clahe	0.111	-0.006
EfficientNet-B3	IDRiD	color transfer	0.127	+0.011
RETFound full	Messidor-2	baseline	0.395	+0.000
RETFound full	Messidor-2	clahe	0.606	+0.211
RETFound full	Messidor-2	color transfer	0.526	+0.131
RETFound full	IDRiD	baseline	0.139	+0.000
RETFound full	IDRiD	clahe	0.293	+0.154
RETFound full	IDRiD	color transfer	0.118	-0.021

Inference only: all models were trained at 224 × 224 and evaluated at 384 × 384; RETFound used a bicubically interpolated positional embedding, so this is an off-distribution test.

QWK by model, dataset and input resolution

Model	Dataset	res	QWK	acc	macro-F1	referable AUC
EfficientNet-B3	APTOS internal	224	0.876	0.803	0.644	0.980
EfficientNet-B3	Messidor-2	224	0.499	0.626	0.360	0.842
EfficientNet-B3	IDRiD	224	0.760	0.497	0.441	0.962
EfficientNet-B3	APTOS internal	384	0.824	0.756	0.535	0.977
EfficientNet-B3	Messidor-2	384	0.425	0.616	0.330	0.866
EfficientNet-B3	IDRiD	384	0.731	0.600	0.435	0.984
RETFound full	APTOS internal	224	0.882	0.783	0.632	0.967
RETFound full	Messidor-2	224	0.487	0.624	0.368	0.817
RETFound full	IDRiD	224	0.744	0.451	0.419	0.885
RETFound full	APTOS internal	384	0.886	0.776	0.636	0.975
RETFound full	Messidor-2	384	0.553	0.643	0.409	0.851
RETFound full	IDRiD	384	0.652	0.413	0.389	0.856

Internal-to-external QWK drop, and change versus 224 × 224

Model	External	res	QWK drop	delta vs 224
EfficientNet-B3	Messidor-2	224	0.377	+0.000
EfficientNet-B3	IDRiD	224	0.116	+0.000
EfficientNet-B3	Messidor-2	384	0.399	+0.022
EfficientNet-B3	IDRiD	384	0.093	-0.024
RETFound full	Messidor-2	224	0.395	+0.000
RETFound full	IDRiD	224	0.139	+0.000
RETFound full	Messidor-2	384	0.333	-0.062
RETFound full	IDRiD	384	0.234	+0.096

S11.6 Post hoc temperature scaling

The temperature was fitted on the APTOS validation split; ECE used 15 equal-width bins of top-1 confidence.

Fitted temperature (per model)

Model	T	reading
EfficientNet-B3 (baseline)	1.523 (bootstrap 95% CI 1.32-1.71)	over-confident → softened (T>1)
ResNet-50 (baseline)	1.160 (bootstrap 95% CI 1.01-1.30)	over-confident → softened (T>1)
RETFound (linear probe)	0.503 (bootstrap 95% CI 0.43-0.58)	under-confident → sharpened (T<1)
RETFound (full fine-tune)	0.616 (bootstrap 95% CI 0.54-0.70)	under-confident → sharpened (T<1)

Model	Dataset	n	ECE raw	ECE scaled	Brier raw	Brier scaled	NLL raw	NLL scaled

ECE before and after temperature scaling, by dataset

Model	Dataset	n	ECE raw	ECE scaled	Brier raw	Brier scaled	NLL raw	NLL scaled
EfficientNet-B3 (baseline)	APTOS internal	733	0.101	0.043	0.300	0.280	0.645	0.549
EfficientNet-B3 (baseline)	Messidor-2	1744	0.279	0.213	0.644	0.594	1.742	1.277
EfficientNet-B3 (baseline)	IDRiD	455	0.294	0.195	0.764	0.692	1.667	1.328
ResNet-50 (baseline)	APTOS internal	733	0.058	0.053	0.311	0.309	0.608	0.593
ResNet-50 (baseline)	Messidor-2	1744	0.265	0.240	0.650	0.633	1.583	1.420
ResNet-50 (baseline)	IDRiD	455	0.260	0.223	0.778	0.749	1.445	1.357
RETFound (linear probe)	APTOS internal	733	0.182	0.045	0.395	0.344	0.815	0.753
RETFound (linear probe)	Messidor-2	1744	0.121	0.100	0.594	0.591	1.181	1.190
RETFound (linear probe)	IDRiD	455	0.127	0.299	0.758	0.874	1.462	1.804
RETFound (full fine-tune)	APTOS internal	733	0.179	0.081	0.369	0.335	0.748	0.721
RETFound (full fine-tune)	Messidor-2	1744	0.095	0.090	0.563	0.558	1.158	1.179
RETFound (full fine-tune)	IDRiD	455	0.129	0.229	0.715	0.778	1.377	1.634

Temperature scaling is monotonic in the logits, so accuracy and QWK are unchanged (verified); only calibration changes.

S11.7 Feature-space out-of-distribution score versus entropy as an error detector

Model	Dataset	n	errors	AUC(OOD)	AUC(entropy)	diff (OOD-entropy), 95% CI	verdict
EfficientNet-B3	APTOS internal	733	144	0.652	0.841	-0.189 (-0.243,-0.136)	entropy better
EfficientNet-B3	Messidor-2	1744	653	0.472	0.681	-0.210 (-0.244,-0.176)	entropy better
EfficientNet-B3	IDRiD	455	229	0.486	0.592	-0.106 (-0.174,-0.032)	entropy better
ResNet-50	APTOS internal	733	155	0.644	0.788	-0.145 (-0.197,-0.091)	entropy better
ResNet-50	Messidor-2	1744	674	0.474	0.652	-0.178 (-0.209,-0.146)	entropy better
ResNet-50	IDRiD	455	251	0.506	0.547	-0.041 (-0.106,+0.018)	no sig. difference
RETFound LP	APTOS internal	733	177	0.663	0.695	-0.032 (-0.101,+0.030)	no sig. difference
RETFound LP	Messidor-2	1744	716	0.479	0.649	-0.170 (-0.206,-0.135)	entropy better
RETFound LP	IDRiD	455	302	0.427	0.648	-0.221 (-0.322,-0.123)	entropy better
RETFound full	APTOS internal	733	159	0.637	0.516	+0.120 (+0.040,+0.197)	OOD better
RETFound full	Messidor-2	1744	655	0.452	0.633	-0.181 (-0.215,-0.144)	entropy better
RETFound full	IDRiD	455	250	0.556	0.491	+0.065 (-0.026,+0.150)	no sig. difference

APTOS internal rows are a control: the out-of-distribution detector is fitted on the same distribution, so an advantage there reflects in-domain behavior rather than detection of dataset shift.

S11.8 Classifier-head depth and activation sweep (frozen features)

MCC, Matthews correlation coefficient.

Features: EfficientNet-B3 and RETFound (ViT-L/16) encoders, evaluated on the APTOS internal set, Messidor-2 and IDRiD. Grid: head depth 0, 3, 4, 5, 6 or 7 layers × activation GELU, identity, ReLU or tanh × 5 seeds. Total: 630 trained heads.

Seed noise floor

Standard deviation across seed replicates within each model–depth–activation cell. A depth effect smaller than this cannot be distinguished from seed variation.

- median within-cell SD of MCC: 0.0169
- maximum within-cell SD of MCC: 0.0603

Mean MCC by depth × activation (averaged over seeds)

Head depth	GELU	Identity	ReLU	Tanh
0	n/a	0.4135	n/a	n/a
3	0.414	0.4236	0.4193	0.4261
4	0.422	0.4174	0.4226	0.423
5	0.4266	0.4085	0.4279	0.4154
6	0.4164	0.4028	0.4132	0.4185
7	0.4196	0.4142	0.4097	0.4247

Mean balanced accuracy by depth × activation

Head depth	GELU	Identity	ReLU	Tanh
0	n/a	0.4947	n/a	n/a
3	0.4839	0.4943	0.4839	0.4961
4	0.4868	0.4932	0.4854	0.4924
5	0.4824	0.4847	0.4864	0.4901
6	0.4686	0.486	0.4744	0.4912
7	0.4751	0.487	0.4779	0.4933

Best cell

- EfficientNet-B3 features, APTOS internal set, depth 4, GELU activation: mean MCC 0.7009

On a grid of this size the top cell may reflect chance rather than a better architecture; it should be read against the seed noise floor above.

Depth effect against the noise floor

- best-minus-worst depth, averaged over activations and seeds: 0.0085
- median within-cell seed SD: 0.0169
- ratio: 0.51x

A ratio below about 2 indicates that head depth changed performance no more than reseeding alone; head depth therefore did not matter for this task.