Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Pravinraj Ethiraj

pravinraje.aids2024@citchennai.net

Chennai Institute of Technology

Research Article

Keywords: Diabetic retinopathy, Deep learning, Foundation model, External validation, Reference standard, Fundus photography

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11702966/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Purpose

To determine whether retinal foundation-model pretraining reduces the loss of diabetic retinopathy (DR) grading agreement on external fundus-photograph datasets compared with ImageNet-pretrained convolutional networks.

Methods

RETFound (linear probe and full fine-tuning), EfficientNet-B3 and ResNet-50 were developed on APTOS 2019 (3,662 images) and tested unchanged on Messidor-2 (1,744 images, adjudicated grades) and IDRiD (516 images). The prespecified primary outcome was the difference in internal-to-external quadratic weighted kappa (QWK) drop relative to EfficientNet-B3 for one training run per model; a post hoc analysis averaged three or four runs per model.

Results

Internal QWK was 0.879 (EfficientNet-B3) and 0.888 (RETFound). In the prespecified analysis, RETFound's excess drop was + 0.020 (95% CI − 0.024 to + 0.064) on Messidor-2 and + 0.033 (− 0.023 to + 0.091) on IDRiD; averaged over runs, it was + 0.068 (+ 0.003 to + 0.134; Holm p = 0.074) and + 0.020 (− 0.041 to + 0.080). On Messidor-2, referable-DR sensitivity was 36–43% with specificity 90–98%, and EfficientNet-B3 graded 259 of 270 mild-DR images as no DR. Matching case mix and merging grades 0 and 1 reduced the Messidor-2 drop by 63–69%. Label smoothing reproduced RETFound's lower external calibration error.

Conclusion

Retinal foundation-model pretraining did not reduce external performance loss. Similar internal agreement did not establish robustness across image sources and reference standards, and no model was suitable for autonomous screening at these operating points.

Introduction

Diabetic retinopathy (DR) remains a leading cause of preventable vision loss, and the number of people who need regular retinal screening is growing faster than the number of graders available to read their photographs; in India, where two of the datasets used here originate, teleophthalmology and AI-assisted screening are being scaled to close this gap [1]. Deep learning systems trained on color fundus photographs can reach specialist-level agreement on the data they were developed with [2, 3], and this journal has published deep-learning approaches to DR lesion analysis [4, 5] and evaluations of

commercial DR software in screening settings [6]. Performance, however, often falls when the camera, population, image quality or grading protocol changes [7, 8]; the same validated algorithms perform differently when fundus photographs of the same patients are taken with a desktop and a portable camera [9]. Reviews of DR screening artificial intelligence (AI) note that external validation is still reported inconsistently [10, 11], and systematic reviews in this journal have found that AI studies in ophthalmology rarely include external validation [12, 13] and that results with limited generalizability should be applied cautiously in real-world settings [14]. The opening editorial of this journal's special collection on multimodal AI in retinal disease asks that new models show incremental value over strong baselines, address distribution shift, use transparent reference standards and independent external validation, assess uncertainty, and report outcomes such as referral appropriateness [15].

Retinal foundation models have been proposed as one route to robustness. RETFound, a Vision Transformer [16] pretrained by masked autoencoding [17] on about 1.6 million unlabeled retinal images, was reported to transfer well to several ocular and systemic tasks [18]. The evidence that this improves transfer is mixed. On the three public DR datasets used here, a natural-image foundation model generally achieved higher cross-dataset AUC than RETFound [19]. In a large comparison, RETFound and ImageNet-pretrained models performed similarly when fine-tuned on full datasets, with RETFound advantages confined to small-data regimes [20]. Frozen RETFound features transferred worse than an ImageNet encoder from APTOS 2019 to Messidor-2 [21], and foundation models generalized poorly to unseen ultra-wide-field populations [22]. Comparisons can also be undermined by the data: in one recent comparison of EfficientNet-B3 and RETFound, the authors' own provenance audit later found the pooled cohort confounded and stated that the original conclusion did not stand [23].

An internal-to-external drop in agreement mixes several things: a change in the images (camera, population, acquisition), a change in case mix (grade prevalence), and a change in the reference standard that defines the labels. Adjudicated specialist grades differ materially from single-grader labels and change the measured performance of the same algorithm [24]. Existing comparisons of retinal foundation models have therefore not separated four things that each can change an external result: retinal pretraining from the fine-tuning recipe that accompanies it; differences between model families from random variation between training runs; true image-distribution shift from differences in case mix and reference standard; and genuine calibration advantages from the probability-shaping effect of training choices such as label smoothing. Two further questions are rarely tested quantitatively: whether a model's uncertainty identifies its errors better than its plain confidence, and whether its saliency falls on expert-annotated lesions more often than a trivial control [5, 25].

We therefore compared RETFound, as a frozen linear probe and fully fine-tuned, with two ImageNet-pretrained convolutional neural networks (CNNs), all developed on one dataset and tested unchanged on two external test datasets that differ in camera, population, case mix and reference standard. Whereas cross-camera studies ask whether a change in acquisition alters the performance of an existing system [9], we asked whether retinal-specific pretraining reduces the loss of performance when a model is transferred to new datasets. The primary hypothesis was that fully fine-tuned RETFound loses less five-

grade agreement than EfficientNet-B3 between internal and external data. Secondary aims were to report referable-DR accuracy in clinical terms; to describe how the drop relates to case mix and to the no-DR/mild boundary; to test whether uncertainty identifies errors and supports referral to a human grader; and to measure saliency localization against expert lesion masks.

Methods

Study design and reporting

This was a retrospective external-validation study using only public, de-identified images. Models were developed on one dataset and evaluated, unchanged, on two external datasets (Fig. 1). Reporting follows TRIPOD + AI [11]; the completed checklist, with a map to the CLAIM 2024 [26] and STARD-AI [27] topics, is provided as Online Resource 2. Changes made to the analysis plan during the study are listed in ‘Changes to the analysis plan’.

Table 1 summarizes the three datasets. **APTOS 2019** [28] contains 3,662 labeled color fundus photographs from Aravind Eye Hospital (India), acquired across clinics and cameras and graded by one clinician per image on a 0–4 scale; no adjudication is described. It was used only for development and split 70/10/20 into training (2,562), validation (367) and internal test (733) sets, stratified by grade (seed 42). Forty-one internal-test images had a byte-identical duplicate in the training or validation split and were excluded from all internal estimates, leaving 692 internal-test images. **Messidor-2** [29, 30] contains 874 examinations (1,748 macula-centered images, one per eye) acquired in France, largely with a Topcon TRC NW6 non-mydriatic camera. We used the adjudicated International Clinical Diabetic Retinopathy (ICDR) grades of three retina specialists [24] and the 1,744 images graded gradable. **IDRiD** [31, 32] is the official Disease-Grading release (516 images: 413 training, 103 test, pooled because neither was used for development), acquired with a Kowa VX-10α camera under mydriasis at an eye clinic in Nanded (India) and graded by experts on an ICDR-based 0–4 scale; the number of graders and any adjudication are not reported.

Table 1
Datasets, reference standards and label harmonization

Role	APTOS 2019	Messidor-2	IDRiD (official)
	Development	**External test**	**External test**
Images used	3,662 (internal test 692[a])	1,744 of 1,748	516
Country / setting	India; Aravind Eye Hospital, multiple clinics	France; Messidor program partners	India; eye clinic, Nanded
Camera	Multiple (not specified)	Topcon TRC NW6[b]	Kowa VX-10α
Mydriasis	Not reported	None[b]	Yes
Native resolution (pixels)	Heterogeneous	1440×960 to 2304×1536	4288×2848
Original grading scale	0–4 clinician scale	ICDR 0–4	ICDR-based 0–4
Reference standard	One clinician; no adjudication described	Three retina specialists; adjudicated [24]	Medical experts; number and adjudication not reported
Eyes per participant	Not identifiable	Two (official examination pairing file)	Not identifiable
Mapping to common scale	Identity (0–4)	Identity (0–4)	Identity (0–4)
Grades 0/1/2/3/4 (%)	49.3/10.1/27.3/5.3/8.1	58.3/15.5/19.9/4.3/2.0	32.6/4.8/32.6/18.0/12.0
Referable DR (grade ≥ 2), %	40.6	26.2	62.6

ICDR, International Clinical Diabetic Retinopathy scale. [a]733 images minus 41 with a byte-identical duplicate in the training or validation split. [b]Messidor-Extension subset; not stated for the Messidor-Original subset.

All three label sets use the same five levels (0 no DR, 1 mild, 2 moderate, 3 severe non-proliferative, 4 proliferative DR), so no relabeling was required; identical numbers do not imply identical grading processes (Table 1). Referable DR was defined as grade ≥ 2. An earlier version of this analysis used a third-party copy of IDRiD; comparison with the official release showed 61 missing images and 27 of 455 labels (5.9%) that differed from the official grade, 11 of them official mild or moderate DR labeled as no DR. All IDRiD results reported here use the official release (audit in Online Resource 3).

Preprocessing

One script with fixed parameters was applied identically to all images of all three datasets. Each image was converted to RGB; the black frame was removed by cropping to the bounding box of pixels whose mean channel intensity exceeded 7 (on a 0–255 scale); the crop was zero-padded to a centered square without distortion and resized to 224 × 224 pixels with bilinear interpolation. This size is RETFound's native input and was used for all models so that pretraining, not input size, differed between families. Images were not rotated or orientation-corrected, no contrast enhancement was applied, and no image was excluded for quality other than the four Messidor-2 images graded ungradable in the reference standard. Pixel values were normalized with the ImageNet channel means (0.485, 0.456, 0.406) and standard deviations (0.229, 0.224, 0.225). Training-time augmentation is described below.

Models and training

EfficientNet-B3 [33] (prespecified primary comparator) and ResNet-50 [34] were initialized from ImageNet weights and fully fine-tuned. RETFound [18] (ViT-Large/16, color-fundus masked-autoencoder weights) was trained as a linear probe (frozen backbone) and fully fine-tuned. The data pipeline, splits, preprocessing, augmentation (flips, ± 15° rotation, mild brightness and contrast jitter), inverse-frequency class-weighted cross-entropy and early stopping on APTOS validation QWK (patience 7) were shared. Each family used its customary fine-tuning recipe (Online Resource 1, Table S1): RETFound received layer-wise learning-rate decay, label smoothing (0.1), learning-rate warm-up and larger weight decay. Two control experiments separated recipe from pretraining: EfficientNet-B3 retrained with label smoothing 0.1 only (label-smoothing control), and a crossed-recipe experiment, added during the study ('Changes to the analysis plan'), in which EfficientNet-B3 was trained with the RETFound recipe and fully fine-tuned RETFound with the CNN recipe (Table S1).

To characterize run-to-run variation, EfficientNet-B3 and the linear probe were each trained with four seeds (42–45) and fully fine-tuned RETFound with three (42–44); only the seed changed between runs. Optimizer settings, learning-rate schedules, stochastic depth, label smoothing, batch sizes and epoch limits for every model and control are given in Table S1.

Ethics statement

This study is a secondary analysis of three publicly available, de-identified fundus-photograph datasets released for research use (APTOS 2019 through a Kaggle research competition; Messidor-2 by ADCIS for research and teaching; IDRiD under a Creative Commons Attribution 4.0 license). The images were collected and released by the original investigators under their own approvals [24, 28, 29, 31]; no new data were collected, the author had no access to identifying information, and no attempt was made to re-identify any individual. The author's institution does not have an institutional ethics committee. Under the Indian Council of Medical Research (ICMR) National Ethical Guidelines for Biomedical and Health

Research Involving Human Participants [35], research on anonymized data or data available in the public domain is classified as less than minimal risk and is a category eligible for exemption from ethics review. No ethics committee approval or formal exemption was therefore obtained for this analysis, which was conducted in accordance with the Declaration of Helsinki.

External testing and leakage controls

Model code and training defaults were fixed in version control before any external image was scored, checkpoints were selected only on APTOS validation QWK, and no threshold, temperature or other parameter was fitted on external data. Byte-level (MD5) and feature-embedding duplicate checks found no image shared between datasets.

Outcomes and statistical analysis

The primary metric was QWK [36], which credits ordinal proximity. The prespecified primary estimand was, for each external dataset, the difference between a model's internal-to-external QWK drop and that of EfficientNet-B3, using one training run per model (seed 42). The family of interest comprised four contrasts (RETFound linear probe and fully fine-tuned, each on Messidor-2 and IDRiD), tested two-sided with Holm adjustment [37]. ResNet-50 contrasts were secondary. Because the external drop varied between training runs ('Changes to the analysis plan'), the same contrasts were also estimated averaged over all runs of each family. This multi-run estimand was introduced after the second RETFound run's external result had been observed; it is therefore reported as an exploratory, post hoc assessment of training-procedure variability, not as a prespecified confirmatory analysis, and both analyses are presented together (Table 2). For the multi-run analysis, intervals came from a two-level bootstrap (2,000 replicates) [38]: in each replicate, training runs were resampled with replacement within each model family, and images (internal and IDRiD) or examinations (Messidor-2) were resampled, with the same image resample used for every model. With three or four runs per family, this estimates run-to-run variation only imprecisely, so intervals conditional on the observed runs are also reported. Throughout, 95% confidence intervals are unadjusted for multiplicity, whereas *p* values for the four RETFound contrasts are Holm-adjusted; an interval that excludes zero can therefore accompany an adjusted *p* value above 0.05. No equivalence margin was prespecified, so non-significant differences are not interpreted as equivalence. The sample size was fixed by the public releases, so precision is conveyed by the width of the confidence intervals rather than by a power calculation.

Table 2

Main comparison: internal-to-external drop in quadratic weighted kappa relative to EfficientNet-B3 under the prespecified single-run analysis and the post hoc multi-run analysis

Model (runs)	Drop, mean (range)	Post hoc: mean over runs Excess drop (95% CI)[a]	Holm *p*	Prespecified: seed-42 run Excess drop (95% CI)[b]	Holm *p*
Messidor-2					
EfficientNet-B3 (4)	0.382 (0.355–0.418)	reference		reference	
RETFound LP (4)	0.467 (0.451–0.491)	+ 0.085 (+ 0.021 to + 0.146)	0.039	+ 0.078 (+ 0.012 to + 0.141)	0.066
RETFound full (3)	0.449 (0.401–0.477)	+ 0.068 (+ 0.003 to + 0.134)	0.074	+ 0.020 (− 0.024 to + 0.064)	0.492
ResNet-50 (1)[c]	0.446	–		+ 0.065 (+ 0.014 to + 0.116)	–
IDRiD					
EfficientNet-B3 (4)	0.173 (0.151–0.210)	reference		reference	
RETFound LP (4)	0.267 (0.256–0.282)	+ 0.094 (+ 0.034 to + 0.155)	0.016	+ 0.110 (+ 0.046 to + 0.175)	0.008
RETFound full (3)	0.193 (0.169–0.224)	+ 0.020 (− 0.041 to + 0.080)	0.550	+ 0.033 (− 0.023 to + 0.091)	0.492
ResNet-50 (1)[c]	0.218	–		+ 0.065 (+ 0.013 to + 0.119)	–

Drop: internal (APTOS, *n* = 692) minus external QWK; a positive excess drop means a larger drop than EfficientNet-B3. 95% CIs are unadjusted for multiplicity; *p* values are Holm-adjusted over the four RETFound contrasts. [a]Exploratory analysis introduced after the second RETFound run was observed; two-level bootstrap resampling training runs within each model family and images or Messidor-2 examinations (2,000 replicates); intervals conditional on the observed runs are in Online Resource 1. [b]Originally specified analysis; seed-42 runs only. [c]Secondary comparator, single run. LP, linear probe.

Messidor-2 contributes two eyes per examination. Examinations were identified from the official pairing file distributed with the dataset (870 examinations with both images gradable and 4 with one). Fellow eyes had identical reference grades in 76% of examinations, against 41% expected for two random images, and the within-examination correlation of EfficientNet-B3 grading errors was 0.55; all Messidor-2 intervals therefore resample whole examinations.

Secondary outcomes were sensitivity, specificity, positive and negative predictive values (PPV, NPV) for referable DR at each model's own grade decision, with 95% bootstrap intervals; accuracy; and referable-DR area under the receiver operating characteristic curve (AUC) from the summed probability of grades 2–4, compared with EfficientNet-B3 by paired bootstrap with Holm adjustment within each dataset. Sensitivity and specificity therefore describe one operating point, the grade the model ranks highest, whereas the AUC summarizes discrimination across all thresholds; no alternative threshold was selected, because a threshold chosen on external data would be biased and none had been prespecified on development data. Calibration was summarized by the expected calibration error (ECE; 15 equal-width bins) [39]. Secondary analyses used the seed-42 run of each model.

Case mix and the no-DR/mild boundary

To describe how the drop depends on case mix, each external image was weighted by the ratio of the APTOS internal-test prevalence of its reference grade to the external prevalence of that grade, so that the weighted external set has the development grade mix. Because graders disagree most at the no-DR/mild boundary [24], QWK was also computed on a four-level scale with grades 0 and 1 merged. The drop was recomputed under each correction and both together. These corrections are descriptive: merging grades 0 and 1 removes both any difference in where graders place that boundary and the model's task of detecting subtle lesions, and the two cannot be separated with these data.

Uncertainty and selective referral

For each image we computed the maximum softmax probability (MSP, a simple confidence baseline) [40], the margin between the two highest probabilities, and the predictive entropy (the uncertainty score in the original analysis plan). Error detection was summarized by the AUC for separating incorrectly from correctly graded images, with paired differences from MSP. For selective referral [41], the least confident 10–30% of images were referred to a human grader; we report the fraction of all referable-DR images that the system would miss (retained and graded non-referable), with a fixed random referral order as a no-information comparator, and accuracy and QWK on retained images.

Saliency localization against expert lesion masks

The IDRiD Segmentation subset (81 images with pixel-level expert masks for microaneurysms, hemorrhages, hard and soft exudates, and the optic disc) [31], previously used for lesion segmentation [4], was used for quantitative localization; none of these images was used for training or model selection. Saliency maps for the predicted grade were computed with Grad-CAM++ [42] on the last convolutional block of the CNNs and with gradient-weighted class activation over the last transformer block's patch tokens for RETFound. Masks were transformed with the same crop, pad and resize as the images. Within the field of view we computed the pixel-level AUC of the saliency map for lesion versus non-lesion pixels, the pointing game (saliency peak within 4 pixels of a lesion) [43], lift (saliency mass on

lesions divided by lesion area share), precision of the top 10% of salient pixels, and saliency mass on the optic disc. Controls were a centered Gaussian prior and each model's map for a different image. Because CNN maps derive from a 7 × 7 grid and RETFound maps from a 14 × 14 grid, RETFound maps were also re-scored after degradation to 7 × 7. Paired comparisons used the Wilcoxon signed-rank test.

Changes to the analysis plan

The original plan specified one training run per model (seed 42), image-level resampling, all 733 internal-test images, and a third-party copy of IDRiD. Five changes were made, each before the final analysis and each reported with the original result. (1) The IDRiD copy was replaced by the official release after the label audit above. (2) Duplicated internal-test images were excluded. (3) Messidor-2 examinations were resampled as clusters using the official pairing file, replacing a worst-case design-effect bound. (4) A second fine-tuning run of RETFound, trained to estimate run-to-run variation, gave a materially larger excess drop than the first, so a third run was trained and the contrasts were re-estimated averaged over runs; this analysis is exploratory and is reported alongside the originally specified one (Table 2). The additional EfficientNet-B3 and linear-probe runs had been completed earlier as a sensitivity analysis. (5) The crossed-recipe control was added to separate pretraining from the fine-tuning recipe.

Software, reproducibility and use of generative AI

Analyses used Python with PyTorch 2.11 and timm 1.0.29 on a single laptop GPU (NVIDIA RTX 5060). Checkpoints of the seed-42 runs are identified by SHA-256 hash (Online Resource 1). All manuscript numbers are generated by script from the per-image prediction files. Large-language-model assistants (Claude, Anthropic; Codex, OpenAI) were used under the author's direction to write and debug code, generate figures, and draft and edit text; all analyses were executed by the author, every number was checked against the prediction files, and the author takes full responsibility for the content.

Results

The main result is the comparison of internal-to-external drops (Table 2, Fig. 2). Referable-DR performance and the case-mix analysis are secondary; the recipe controls, calibration, uncertainty and saliency analyses are exploratory and are reported after them.

Internal performance

On the deduplicated APTOS internal test set (n = 692), QWK was 0.879 (95% CI 0.850–0.905) for EfficientNet-B3, 0.880 (0.853–0.904) for ResNet-50, 0.852 (0.823–0.880) for the RETFound linear probe and 0.888 (0.863–0.910) for fully fine-tuned RETFound (Table 3). Referable-DR sensitivity was 85–90%, specificity 93–96% and AUC 0.964–0.979; no model's AUC differed from EfficientNet-B3's after adjustment (Holm $p \geq 0.276$). Including the 41 duplicated images changed internal QWK by at most 0.007.

Table 3
Grading agreement and referable diabetic retinopathy performance on internal and external data (seed-42 run of each model)

Model	QWK (95% CI)	Sensitivity, % (95% CI)	Specificity, % (95% CI)	PPV, %	NPV, %	AUC (95% CI)
APTOS internal (n = 692)						
EfficientNet-B3	0.879 (0.850–0.905)	90 (86–93)	93 (91–96)	89	93	0.979 (0.970–0.987)
ResNet-50	0.880 (0.853–0.904)	87 (82–90)	96 (93–97)	92	92	0.978 (0.968–0.986)
RETFound LP	0.852 (0.823–0.880)	85 (80–89)	96 (93–97)	92	91	0.976 (0.964–0.986)
RETFound full	0.888 (0.863–0.910)	90 (86–93)	96 (93–97)	93	94	0.964 (0.944–0.981)
Messidor-2 (n = 1,744)						
EfficientNet-B3	0.499 (0.434–0.557)	39 (33–44)	97 (96–98)	84	82	0.842 (0.816–0.867)
ResNet-50	0.435 (0.369–0.495)	36 (30–41)	96 (94–97)	74	81	0.789 (0.758–0.818)
RETFound LP	0.394 (0.321–0.460)	43 (37–49)	90 (88–92)	61	82	0.781 (0.750–0.812)
RETFound full	0.487 (0.425–0.546)	39 (34–45)	98 (97–98)	85	82	0.817 (0.787–0.845)
IDRiD (n = 516)						
EfficientNet-B3	0.726 (0.681–0.768)	83 (78–87)	91 (86–95)	94	76	0.927 (0.905–0.948)
ResNet-50	0.662 (0.604–0.711)	72 (67–77)	95 (92–98)	96	67	0.912 (0.887–0.937)

Seed-42 run of each model. Sensitivity, specificity, PPV and NPV refer to referable diabetic retinopathy (grade ≥ 2) at the model's own grade decision, with 95% bootstrap intervals; AUC is the area under the receiver operating characteristic curve for referable DR. Messidor-2 intervals resample examinations. QWK, quadratic weighted kappa; PPV, NPV, positive and negative predictive value; LP, linear probe.

Model	QWK (95% CI)	Sensitivity, % (95% CI)	Specificity, % (95% CI)	PPV, %	NPV, %	AUC (95% CI)
RETFound LP	0.590 (0.533–0.643)	83 (79–87)	66 (59–72)	80	69	0.882 (0.853–0.911)
RETFound full	0.703 (0.653–0.747)	73 (69–78)	93 (89–96)	94	68	0.876 (0.846–0.907)

Seed-42 run of each model. Sensitivity, specificity, PPV and NPV refer to referable diabetic retinopathy (grade ≥ 2) at the model's own grade decision, with 95% bootstrap intervals; AUC is the area under the receiver operating characteristic curve for referable DR. Messidor-2 intervals resample examinations. QWK, quadratic weighted kappa; PPV, NPV, positive and negative predictive value; LP, linear probe.

External performance and the main comparison

Every model degraded on both external sets, more on Messidor-2 than on IDRiD (Table 3, Fig. 2). In the prespecified single-run analysis, fine-tuned RETFound's excess drop relative to EfficientNet-B3 was + 0.020 (unadjusted 95% CI − 0.024 to + 0.064; Holm p = 0.492) on Messidor-2 and + 0.033 (− 0.023 to + 0.091; Holm p = 0.492) on IDRiD, and the linear probe's was + 0.078 (Holm p = 0.066) and + 0.110 (Holm p = 0.008) (Table 2).

In the post hoc multi-run analysis, EfficientNet-B3's QWK fell on average by 0.382 on Messidor-2 and 0.173 on IDRiD, and fully fine-tuned RETFound's by 0.449 and 0.193 (Table 2). Fine-tuned RETFound's excess drop relative to EfficientNet-B3 was + 0.068 (unadjusted 95% CI + 0.003 to + 0.134; Holm p = 0.074) on Messidor-2 and + 0.020 (− 0.041 to + 0.080; Holm p = 0.550) on IDRiD. Conditional on the observed runs, the intervals were + 0.027 to + 0.108 and − 0.026 to + 0.065. The Messidor-2 interval excludes zero but is unadjusted for multiplicity; after Holm adjustment the difference was not statistically significant. Neither analysis showed a smaller external drop for RETFound: on Messidor-2 the point estimate was larger for RETFound in both, and on IDRiD the two were similar. Neither result establishes equivalence. The drop varied between training runs of the same model: across RETFound runs it ranged from 0.401 to 0.477 on Messidor-2, and across EfficientNet-B3 runs from 0.355 to 0.418. On Messidor-2 the first RETFound run therefore showed a smaller difference than the mean across runs.

Averaged over runs, the linear probe degraded more than EfficientNet-B3 on both sets (+ 0.085, + 0.021 to + 0.146, Holm p = 0.039; +0.094, + 0.034 to + 0.155, Holm p = 0.016). In the secondary comparison, ResNet-50 also degraded more (+ 0.065 and + 0.065).

Secondary: referable DR in clinical terms

On Messidor-2, specificity for referable DR was preserved (90–98%) but sensitivity was low (36–43%; Table 3). Of 270 reference-standard mild images, EfficientNet-B3 graded 259 as no DR and fine-tuned RETFound 266; of 347 moderate images, 249 and 263, respectively, were graded no DR (Fig. 3b). On IDRiD, sensitivity was higher (72–83%) and errors went the other way at the same boundary: of 168

reference-standard no-DR images, EfficientNet-B3 graded 82 as mild and RETFound 88. Referable-DR AUC was 0.781–0.842 on Messidor-2 and 0.876–0.927 on IDRiD; EfficientNet-B3 had a higher AUC than fine-tuned RETFound on IDRiD (difference − 0.051, 95% CI − 0.081 to − 0.019; Holm p = 0.006) but not on Messidor-2 (− 0.025; Holm p = 0.165).

These sensitivities describe a single operating point, each model's highest-ranked grade, and are not suitable for autonomous screening, where missing more than half of referable DR would be unacceptable. Ordinal grading agreement and binary referable-DR discrimination are distinct: the referable-DR AUC of 0.842 for EfficientNet-B3 on Messidor-2 indicates that the models still ranked referable above non-referable images reasonably well, so a lower decision threshold could trade specificity for sensitivity. No such threshold was prespecified.

Secondary: case mix and the no-DR/mild boundary

This analysis describes how the observed drop changes under each correction; it does not identify the cause of the drop. Matching the external case mix to APTOS reduced the Messidor-2 drop by 23% (EfficientNet-B3) and 28% (RETFound); merging grades 0 and 1 reduced it by 47% and 48%; together by 63% and 69% (Fig. 3a; all models in Online Resource 1, S8). The remaining drops (0.143, 95% CI 0.047–0.249; 0.124, 0.045–0.205) excluded zero, so the corrections did not account for all of the loss. On IDRiD, reweighting alone did not reduce the drop, because IDRiD's pathology-enriched case mix favors QWK; the two corrections together reduced EfficientNet-B3's drop by 69% and RETFound's by 35%. After both corrections, fine-tuned RETFound's excess drop was − 0.019 (− 0.103 to + 0.052) on Messidor-2 and + 0.072 (− 0.006 to + 0.155) on IDRiD.

Exploratory: matched-recipe and label-smoothing controls

In the crossed-recipe experiment (single seed-42 runs), RETFound had no advantage under either recipe. When both families were trained with the RETFound recipe, RETFound's excess drop relative to EfficientNet-B3 was + 0.021 (95% CI − 0.023 to + 0.067) on Messidor-2 and + 0.010 (− 0.044 to + 0.068) on IDRiD; when both were trained with the CNN recipe, it was + 0.088 (+ 0.043 to + 0.132) and − 0.006 (− 0.063 to + 0.047). The label-smoothed EfficientNet-B3 control did not differ from the original EfficientNet-B3 in drop (+ 0.029, − 0.010 to + 0.071; +0.039, 0.000 to + 0.080).

Exploratory: calibration

Externally, the CNNs were over-confident (ECE 0.279 and 0.320 for EfficientNet-B3 on Messidor-2 and IDRiD), whereas fine-tuned RETFound had lower external ECE (0.095 and 0.170); internally the pattern reversed (0.094 vs 0.192). The label-smoothed EfficientNet-B3 reproduced the RETFound pattern (internal ECE 0.207; Messidor-2 0.041; IDRiD 0.185), as did EfficientNet-B3 trained with the full RETFound recipe (Messidor-2 ECE 0.048). Conversely, RETFound trained with the CNN recipe was well calibrated internally (ECE 0.049) and over-confident externally (Messidor-2 0.245; IDRiD 0.156), like the CNNs. These findings are consistent with the calibration difference arising from the training recipe rather than from retinal pretraining, although the crossed-recipe controls were single runs.

Exploratory: uncertainty and referral

Predictive entropy did not identify grading errors significantly better than MSP for any model (Fig. 4c). For EfficientNet-B3, MSP error-detection AUC was 0.850 internally, 0.682 on Messidor-2 and 0.585 on IDRiD, and the entropy–MSP differences were 0.000, 0.000 and + 0.001. For fine-tuned RETFound, entropy was worse than MSP internally (− 0.123, 95% CI − 0.160 to − 0.089), as for the label-smoothed CNN (− 0.154). On IDRiD, MSP detected fine-tuned RETFound's errors no better than chance (AUC 0.518, 0.469–0.568).

When the least confident 30% of images were referred to a human grader, the fraction of all referable DR missed by the system on Messidor-2 fell from 61% to 28% for EfficientNet-B3 and from 61% to 26% for RETFound, compared with 42% for random referral (Fig. 4a). On IDRiD the reduction was small and no better than random referral (EfficientNet-B3 17% to 13%, random 13%; RETFound 27% to 19%; Fig. 4b). Accuracy on retained Messidor-2 images rose from 63% to 71% for EfficientNet-B3, but QWK on retained images fell from 0.499 to 0.188, because referral preferentially removed diseased images.

Exploratory: saliency localization against expert lesion masks

Lesions occupied 7.1% of the field of view on average. Fine-tuned RETFound's saliency separated lesion from non-lesion pixels better than EfficientNet-B3's (pixel AUC 0.668, 95% CI 0.634–0.701, vs 0.606, 0.582–0.632; $p < 0.001$), placed its peak on a lesion more often (0.64 vs 0.48) and put less mass on the optic disc (Fig. 5). EfficientNet-B3's maps were not significantly better than the centered-Gaussian control (pixel AUC 0.606 vs 0.587, $p = 0.082$), although all models' maps beat the cross-image control ($p < 0.001$), so they tracked image-specific content. Part of RETFound's advantage reflected its finer grid: at a matched 7 × 7 resolution its lift over EfficientNet-B3 was no longer significant (1.440 vs 1.319, $p = 0.174$), whereas pixel AUC (0.661 vs 0.606, $p < 0.001$) and top-10% precision (0.157 vs 0.120, $p = 0.003$) remained higher. In absolute terms localization was weak for every model: at most 19% of the most salient pixels fell on lesions.

Additional analyses

Monte Carlo dropout [44], temperature scaling, alternative loss functions, re-fitted linear heads and inference-time contrast enhancement and color transfer are reported in Online Resource 1; these were run with the earlier IDRiD copy, and their IDRiD results are given for completeness only.

Discussion

In this external validation, retinal foundation-model pretraining did not reduce the loss of DR grading agreement on two external test datasets. Fully fine-tuned RETFound and EfficientNet-B3 were similar internally. Externally, neither the prespecified single-run analysis nor the post hoc multi-run analysis showed a smaller drop for RETFound: its drop was larger on Messidor-2 in both (not statistically

significant after adjustment) and similar on IDRiD. A frozen RETFound linear probe degraded more than the CNN. Because no equivalence margin was prespecified, these results show the absence of a detectable advantage, not that the two approaches are equivalent. Neither fine-tuning recipe gave RETFound an advantage when the two families were trained with each other's recipes. These findings agree with reports that fine-tuned RETFound performs similarly to ImageNet models on full datasets [20], that a natural-image foundation model matches or exceeds it across the same three datasets [19], and that frozen RETFound features transfer worse from APTOS to Messidor-2 [21]. In the terms of the opening editorial of this journal's collection on multimodal retinal AI [15], we found no incremental robustness over a strong conventional baseline. The key clinical lesson is that an improvement, or parity, in internal grading agreement does not establish robustness across different image sources and reference standards.

The variability between training runs is itself a finding. The first RETFound run suggested a Messidor-2 drop similar to the CNN's; the next two runs, differing only in seed, degraded more. Single-run comparisons of external performance, which are the norm in this field, can therefore be misleading, and comparisons of model families should report repeated runs, ideally planned in advance and in larger numbers than were available here. The prespecified single-run and the exploratory multi-run analyses led to the same overall conclusion.

The clinically relevant pattern concerns what the drop contains. On Messidor-2 every model preserved specificity, and EfficientNet-B3 and fine-tuned RETFound graded almost all reference-standard mild DR, and about three-quarters of moderate DR, as no DR. At the models' default grade decisions, referable-DR sensitivity was far below what autonomous screening requires, so none of these systems, at these operating points, would be suitable for unsupervised use; these findings are not a basis for clinical deployment. On IDRiD the errors went the other way at the same boundary. In a descriptive decomposition, matching case mix and merging no and mild DR reduced most of the observed Messidor-2 drop. This pattern is consistent with a difference between the single-grader development labels and the adjudicated Messidor-2 grades in where the no-DR/mild boundary is placed [24], but merging grades also removes the hardest detection task, and poorer visibility of subtle lesions in Messidor-2 images, many acquired without mydriasis, could produce the same pattern; our data cannot separate the two, and the decomposition should not be read as showing that grading conventions caused the drop. For ophthalmologists reading AI validation studies, an external drop in an agreement statistic should be read alongside the reference standard of each dataset, the case mix, and prevalence-independent measures such as sensitivity and specificity, as reported in clinical evaluations of DR software [6, 9].

Two commonly cited advantages of RETFound were not shown to come from retinal pretraining. Its lower external ECE was reproduced by adding label smoothing to the CNN and lost when RETFound was trained without its usual recipe; with single control runs, this is consistent with a recipe-related explanation rather than proof of one, and agrees with the known effect of label smoothing on predicted probabilities [45], and the same smoothing may explain why predictive entropy was uninformative:

entropy summarized the deliberately flattened tail of the distribution, whereas the maximum probability still carried signal. Uncertainty estimates should be compared with the model's plain confidence before being presented as added value. In this retrospective selective-referral analysis, referring the least confident 30% of images more than halved the referable DR missed on Messidor-2, a useful but incomplete gain, and gave no benefit over random referral on IDRiD; it is not a screening solution, and its benefit did not generalize across datasets.

Saliency localization, tested against expert lesion masks, favored RETFound, partly because of its finer grid. This does not establish clinical interpretability: every model's maps localized lesions weakly, and the CNN maps were barely distinguishable from a centered prior. Better localization also did not translate into better grading transfer.

Strengths and limitations

Strengths include external datasets never used for any modeling decision, repeated training runs, examination-level resampling for Messidor-2, crossed-recipe and label-smoothing controls, an audit against the official dataset release, and release of code and per-image predictions. This is a retrospective study of curated public datasets; neither external test dataset is a prospective or consecutive screening cohort, so the results describe transfer between research datasets rather than performance at a clinical site. Only one retinal foundation model, one development dataset and one training resolution (224 × 224) were examined, and newer retinal foundation models were not evaluated; downsampling may suppress microaneurysms, and higher resolutions may favor subtle-lesion detection differently for the two families. An inference-only check at 384 pixels with models trained at 224 pixels gave mixed changes (Online Resource 1, S11.5), so a comparison trained at higher resolution remains needed. The multi-run analysis is exploratory. Three or four training runs per family estimate run-to-run variation imprecisely (five to ten would be preferable), and the crossed-recipe controls used single runs. Sensitivity and specificity were reported only at each model's default grade decision; operating points chosen on development data were not evaluated. APTOS and IDRiD provide no patient identifiers, so images from the same person may be correlated, and internal estimates may be optimistic. Image quality was not graded beyond the reference-standard gradability label, and no demographic data were available, so performance could not be examined across age, sex or ethnic subgroups. The corrections for case mix and the no-DR/mild boundary are descriptive and cannot distinguish a grading-convention difference from poorer visibility of subtle lesions. Saliency was evaluated on 81 images from one dataset with one attribution method per family.

Conclusions

In both the prespecified and the post hoc analyses, retinal foundation-model pretraining did not make DR grading more robust to dataset change than an ImageNet-pretrained CNN. External performance varied between training runs and changed markedly with case mix and at the no-DR/mild boundary. External

validation of DR grading systems should report the reference standard of each dataset, prevalence-independent measures and repeated training runs before drawing conclusions about model robustness.

Declarations

Competing interests

The author has no relevant financial or non-financial interests to disclose.

Ethics approval

This study is a secondary analysis of three publicly available, de-identified fundus-photograph datasets released for research use (APTOS 2019 through a Kaggle research competition; Messidor-2 by ADCIS for research and teaching; IDRiD under a Creative Commons Attribution 4.0 license). The images were collected and released by the original investigators under their own approvals [24, 28, 29, 31]; no new data were collected, the author had no access to identifying information, and no attempt was made to re-identify any individual. The author's institution does not have an institutional ethics committee. Under the ICMR National Ethical Guidelines (2017) [35], research on anonymized data or data in the public domain is classified as less than minimal risk and is eligible for exemption from ethics review; no ethics committee approval or formal exemption was therefore obtained. The study was conducted in accordance with the 1964 Declaration of Helsinki and its later amendments.

Consent to participate

Not applicable (secondary analysis of public, de-identified datasets).

Consent to publish

Not applicable. Figure 5 shows de-identified fundus photographs from the IDRiD dataset, released under a Creative Commons Attribution 4.0 license.

Funding

The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

Author Contribution

Pravinraj Ethiraj , conceived and designed the study, curated the data, developed the software, performed the analyses, prepared the figures, and wrote and revised the manuscript. The author read and approved the final manuscript.

Acknowledgments

The author thanks the teams who released APTOS 2019, Messidor-2 and IDRiD, the authors of the adjudicated Messidor-2 grades, and the developers of RETFound for releasing their pretrained weights.

Data Availability

The datasets analysed in this study are publicly available from the following sources: APTOS 2019 through the Kaggle Blindness Detection competition (https://www.kaggle.com/competitions/aptos2019-blindness-detection); Messidor-2 through ADCIS (https://www.adcis.net/en/third-party/messidor2/), with adjudicated grades obtained from Krause et al.; and IDRiD through IEEE DataPort (https://doi.org/10.21227/H25W98). Per-image predictions for all models, training runs and datasets, APTOS split lists, and the IDRiD label audit are provided in Online Resource 3. Analysis code is provided in Online Resource 4.

Code availability

The data-processing, training, evaluation and statistical code, with the software environment and instructions to regenerate every number, table and figure from the prediction files, is provided as Online Resource 4. Trained weights are available from the author on reasonable request.

References

1. Sharma K, Shah K, Raman R (2026) Teleophthalmology for retinal disease screening and referral in low- and middle-income countries: A case overview of India. Int Ophthalmol 46: https://doi.org/10.1007/s10792-026-04171-y
2. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J et al (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 316:2402–2410. https://doi.org/10.1001/jama.2016.17216
3. Ting DSW, Cheung CY, Lim G, Tan GSW, Quang ND, Gan A, Hamzah H, Garcia-Franco R, San Yeo IY, Lee SY et al (2017) Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA 318:2211–2223. https://doi.org/10.1001/jama.2017.18152

4. Yuan H, Dai M, Shi C, Li M, Li H (2023) A generative adversarial neural network with multi-attention feature extraction for fundus lesion segmentation. Int Ophthalmol 43:5079–5090. https://doi.org/10.1007/s10792-023-02911-y
5. Mahapadi AA, Shirsath V, Pundge A (2025) Non-proliferative diabetic retinopathy detection using Rosmarus Quagga optimized explainable generative meta learning based deep convolutional neural network model. Int Ophthalmol 45:178. https://doi.org/10.1007/s10792-025-03528-z
6. Vought R, Vought V, Shah M, Szirth B, Bhagat N (2023) EyeArt artificial intelligence analysis of diabetic retinopathy in retinal screening events. Int Ophthalmol 43:4851–4859. https://doi.org/10.1007/s10792-023-02887-9
7. Beede E, Baylor E, Hersch F, Iurchenko A, Wilcox L, Ruamviboonsuk P, Vardoulakis LM (2020) A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: Proceedings of the 2020 CHI conference on human factors in computing systems. pp 1–12
8. Dai L, Wu L, Li H, Cai C, Wu Q, Kong H, Liu R, Wang X, Hou X, Liu Y et al (2021) A deep learning system for detecting diabetic retinopathy across the disease spectrum. Nat Commun 12:3242. https://doi.org/10.1038/s41467-021-23458-5
9. He S, Bulloch G, Zhang L, Xie Y, Wu W, He Y, Meng W, Shi D, He M (2023) Cross-camera performance of deep learning algorithms to diagnose common ophthalmic diseases: A comparative study highlighting feasibility to portable fundus camera use. Curr Eye Res 48:857–863. https://doi.org/10.1080/02713683.2023.2215984
10. Grzybowski A, Brona P, Lim G, Ruamviboonsuk P, Tan GSW, Abramoff M, Ting DSW (2020) Artificial intelligence for diabetic retinopathy screening: A review. Eye (Lond) 34:451–460. https://doi.org/10.1038/s41433-019-0566-0
11. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, Ghassemi M, Liu X, Reitsma JB, van Smeden M et al (2024) TRIPOD + AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 385:e078378. https://doi.org/10.1136/bmj-2023-078378
12. Kim HK, Yoo TK (2025) Oculomics approaches using retinal imaging to predict mental health disorders: A systematic review and meta-analysis. Int Ophthalmol 45:111. https://doi.org/10.1007/s10792-025-03500-x
13. Hashemi H, Jamali A, Nabovati P, Babapour Mofrad F, Khabazkhoob M (2025) The application of artificial intelligence-based algorithms in predicting the progression of keratoconus: A systematic review. Int Ophthalmol 45:482. https://doi.org/10.1007/s10792-025-03855-1
14. Heidari Z, Mirghorbani M, Abounoori M, Ebrahimibesheli K, Tabarestani M, Khabazkhoob M, Yousefi S, Modjtahedi BS (2026) Vitreoretinal disease detection using artificial intelligence: A systematic review and meta-analysis. Int Ophthalmol 46:77. https://doi.org/10.1007/s10792-026-03960-9
15. Jin K, Grzybowski A (2026) Multimodal AI in retinal disease: From images to decisions and real-world endpoints. Int Ophthalmol 46:378. https://doi.org/10.1007/s10792-026-04254-w

16. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S et al (2021) An image is worth 16x16 words: Transformers for image recognition at scale. In: International conference on learning representations (ICLR)
17. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R (2022) Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp 16000–16009
18. Zhou Y, Chia MA, Wagner SK, Ayhan MS, Williamson DJ, Struyven RR, Liu T, Xu M, Lozano MG, Woodward-Court P et al (2023) A foundation model for generalizable disease detection from retinal images. Nature 622:156–163. https://doi.org/10.1038/s41586-023-06555-x
19. Hou Q, Zhou Y, Lin Goh JH, Zou K, Yew SME, Srinivasan S, Wang M, Lo TWS, Lei X, Wagner SK et al (2025) Can a natural image-based foundation model outperform a retina-specific model in detecting ocular and systemic diseases? Ophthalmol Sci 6:100923. https://doi.org/10.1016/j.xops.2025.100923
20. Yew SME, Lei X, Goh JHL, Zou K, Chen Y, Srinivasan S, Chee ML, Pushpanathan K, Hou Q, Soh ZD et al (2026) Performance and label efficiency of traditional deep-learning models and a retina-specific foundation model for ocular and systemic disease detection: A retrospective comparative study. Lancet Digit Health 8:101031. https://doi.org/10.1016/j.landig.2026.101031
21. Poyrazer M, Yağcı H, Erten R (2026) How well do frozen foundation models transfer? A calibration-focused benchmark for diabetic retinopathy grading. Front Med (Lausanne) 13:1815982. https://doi.org/10.3389/fmed.2026.1815982
22. Li LY, Thambawita V, Byberg S, Hulman A (2026) Assessing the generalisability of foundation models to ultra-wide field retinal imaging for diabetic retinopathy screening in Denmark and Greenland. Int J Med Inform 217:106503. https://doi.org/10.1016/j.ijmedinf.2026.106503
23. Alutaibi M, Alazmi M (2026) Comparative evaluation of EfficientNet-B3, vision transformers, and RETFound for seven-category retinopathy fundus image classification. Diagnostics (Basel) 16:3109. https://doi.org/10.3390/diagnostics16193109
24. Krause J, Gulshan V, Rahimy E, Karth P, Widner K, Corrado GS, Peng L, Webster DR (2018) Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology 125:1264–1272. https://doi.org/10.1016/j.ophtha.2018.01.034
25. Alizadeh Eghtedar R, Vard A, Malekahmadi M, Peyman A (2024) A new computer-aided diagnosis tool based on deep learning methods for automatic detection of retinal disorders from OCT images. Int Ophthalmol 44:110. https://doi.org/10.1007/s10792-024-03033-9
26. Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, Kahn CE (2024) Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. Radiol Artif Intell 6:e240300. https://doi.org/10.1148/ryai.240300
27. Sounderajah V, Guni A, Liu X, Collins GS, Karthikesalingam A, Markar SR, Golub RM, Denniston AK, Shetty S, Moher D et al (2025) The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. Nat Med 31:3283–3289. https://doi.org/10.1038/s41591-025-03953-8

28. Asia Pacific Tele-Ophthalmology Society(2019) APTOS 2019 blindness detection: Kaggle competition dataset. https://www.kaggle.com/competitions/aptos2019-blindness-detection. Accessed 8 Sep 2026
29. Decencière E, Zhang X, Cazuguel G, Lay B, Cochener B, Trone C, Gain P, Ordonez R, Massin P, Erginay A et al (2014) Feedback on a publicly distributed image database: The Messidor database. Image Anal Stereol 33:231–234. https://doi.org/10.5566/ias.1155
30. Abràmoff MD, Folk JC, Han DP, Walker JD, Williams DF, Russell SR, Massin P, Cochener B, Gain P, Tang L et al (2013) Automated analysis of retinal images for detection of referable diabetic retinopathy. JAMA Ophthalmol 131:351–357. https://doi.org/10.1001/jamaophthalmol.2013.1743
31. Porwal P, Pachade S, Kamble R, Kokare M, Deshmukh G, Sahasrabuddhe V, Meriaudeau F (2018) Indian diabetic retinopathy image dataset (IDRiD): A database for diabetic retinopathy screening research. Data 3:25. https://doi.org/10.3390/data3030025
32. Porwal P, Pachade S, Kokare M, Deshmukh G, Son J, Bae W, Liu L, Wang J, Liu X, Gao L et al (2020) IDRiD: Diabetic retinopathy – segmentation and grading challenge. Med Image Anal 59:101561. https://doi.org/10.1016/j.media.2019.101561
33. Tan M, Le QV (2019) EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th international conference on machine learning (ICML), PMLR 97. pp 6105–6114
34. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp 770–778
35. Indian Council of Medical Research(2017) National ethical guidelines for biomedical and health research involving human participants. Indian Council of Medical Research, New Delhi
36. Cohen J (1968) Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 70:213–220. https://doi.org/10.1037/h0026256
37. Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6:65–70
38. Efron B, Tibshirani RJ (1994) An introduction to the bootstrap. Chapman & Hall/CRC, New York
39. Guo C, Pleiss G, Sun Y, Weinberger KQ (2017) On calibration of modern neural networks. In: Proceedings of the 34th international conference on machine learning (ICML), PMLR 70. pp 1321–1330
40. Hendrycks D, Gimpel K (2017) A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International conference on learning representations (ICLR)
41. Geifman Y, El-Yaniv R (2017) Selective classification for deep neural networks. In: Advances in neural information processing systems 30 (NeurIPS)
42. Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN (2018) Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE winter conference on applications of computer vision (WACV). pp 839–847

43. Zhang J, Bargal SA, Lin Z, Brandt J, Shen X, Sclaroff S (2018) Top-down neural attention by excitation backprop. Int J Comput Vis 126:1084–1102. https://doi.org/10.1007/s11263-017-1059-x
44. Gal Y, Ghahramani Z (2016) Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: Proceedings of the 33rd international conference on machine learning (ICML), PMLR 48. pp 1050–1059
45. Müller R, Kornblith S, Hinton G (2019) When does label smoothing help?In: Advances in neural information processing systems 32 (NeurIPS)

Figures

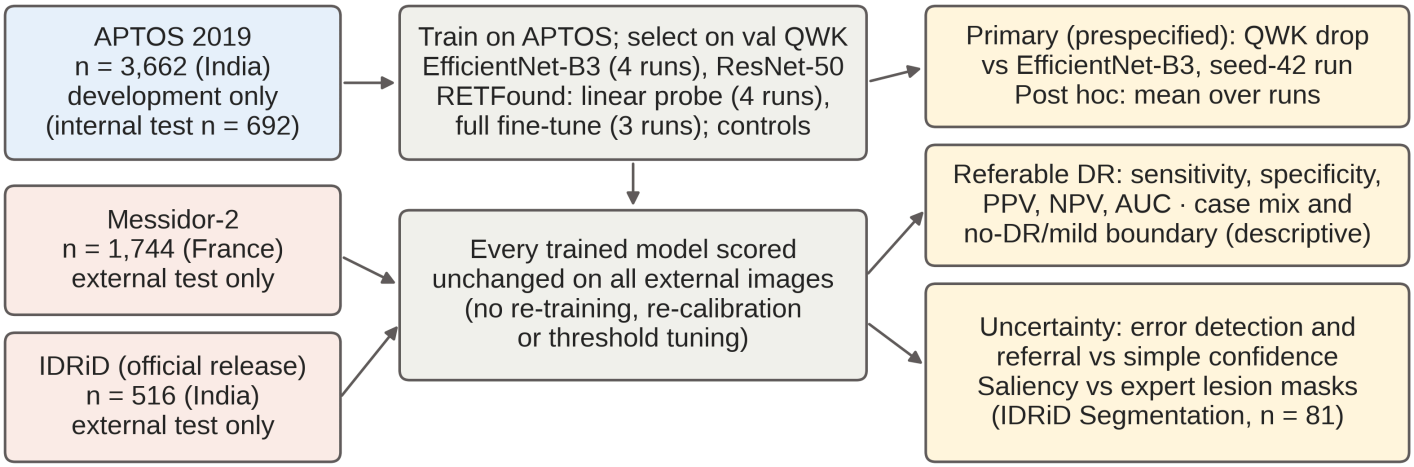

Figure 1

Study design. APTOS 2019 was used only for development; Messidor-2 and IDRiD were held out and scored unchanged for every model. QWK, quadratic weighted kappa

Table 1 summarizes the three datasets. **APTOS 2019** [28] contains 3,662 labeled color fundus photographs from Aravind Eye Hospital (India), acquired across clinics and cameras and graded by one clinician per image on a 0–4 scale; no adjudication is described. It was used only for development and split 70/10/20 into training (2,562), validation (367) and internal test (733) sets, stratified by grade (seed 42). Forty-one internal-test images had a byte-identical duplicate in the training or validation split and were excluded from all internal estimates, leaving 692 internal-test images. **Messidor-2**[29, 30] contains 874 examinations (1,748 macula-centered images, one per eye) acquired in France, largely with a Topcon TRC NW6 non-mydriatic camera. We used the adjudicated International Clinical Diabetic Retinopathy (ICDR) grades of three retina specialists [24] and the 1,744 images graded gradable. **IDRiD**[31, 32] is the official Disease-Grading release (516 images: 413 training, 103 test, pooled because neither was used for development), acquired with a Kowa VX-10α camera under mydriasis at an eye clinic in Nanded

(India) and graded by experts on an ICDR-based 0–4 scale; the number of graders and any adjudication are not reported.

All three label sets use the same five levels (0 no DR, 1 mild, 2 moderate, 3 severe non-proliferative, 4 proliferative DR), so no relabeling was required; identical numbers do not imply identical grading processes (Table 1). Referable DR was defined as grade ≥2. An earlier version of this analysis used a third-party copy of IDRiD; comparison with the official release showed 61 missing images and 27 of 455 labels (5.9%) that differed from the official grade, 11 of them official mild or moderate DR labeled as no DR. All IDRiD results reported here use the official release (audit in Online Resource 3).

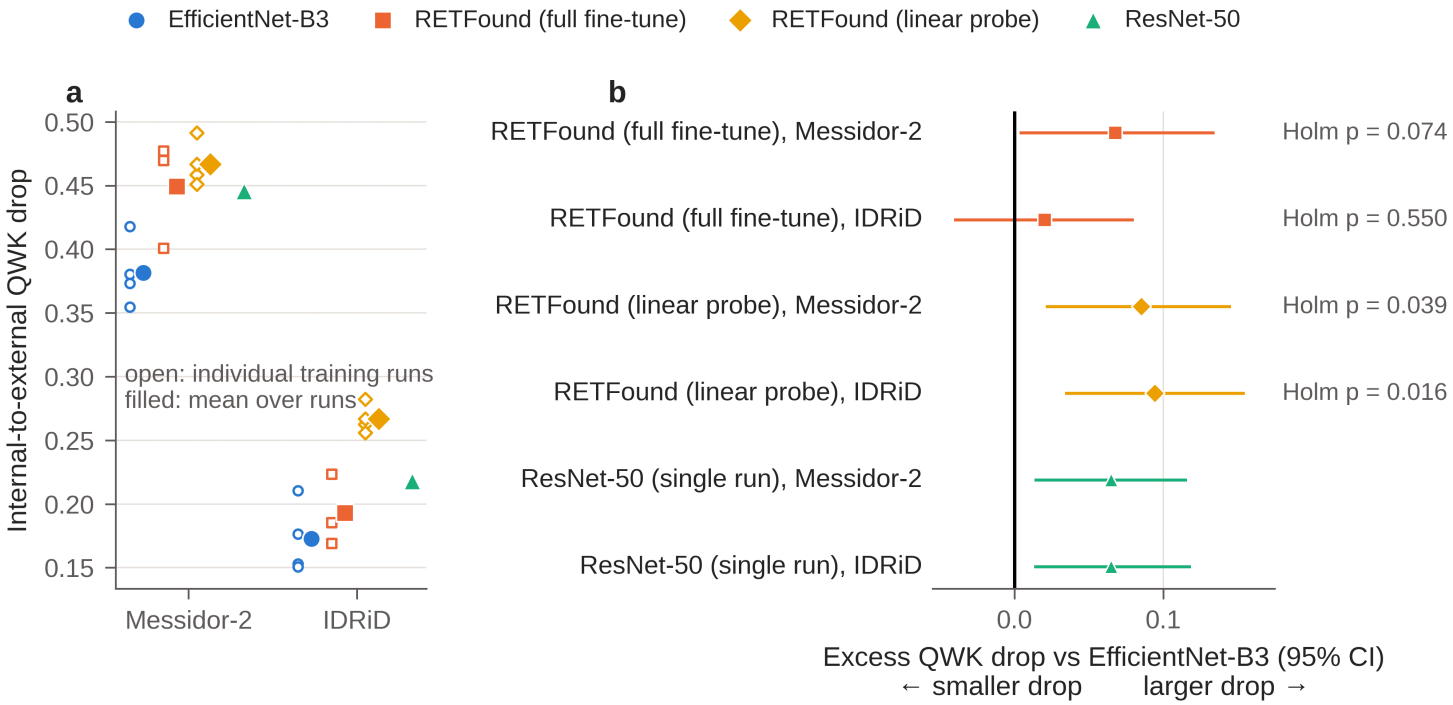

Figure 2

Internal and external grading agreement. **a**Internal-to-external drop in quadratic weighted kappa (QWK) for each training run (small markers) and the mean over runs (large markers) on Messidor-2 and IDRiD. **b** Excess drop relative to EfficientNet-B3 averaged over runs (post hoc analysis), with unadjusted 95% two-level bootstrap intervals and Holm-adjusted *p* values; positive values indicate a larger drop than EfficientNet-B3. ResNet-50 (single run) is secondary. The prespecified single-run (seed-42) contrasts are given in Table 2

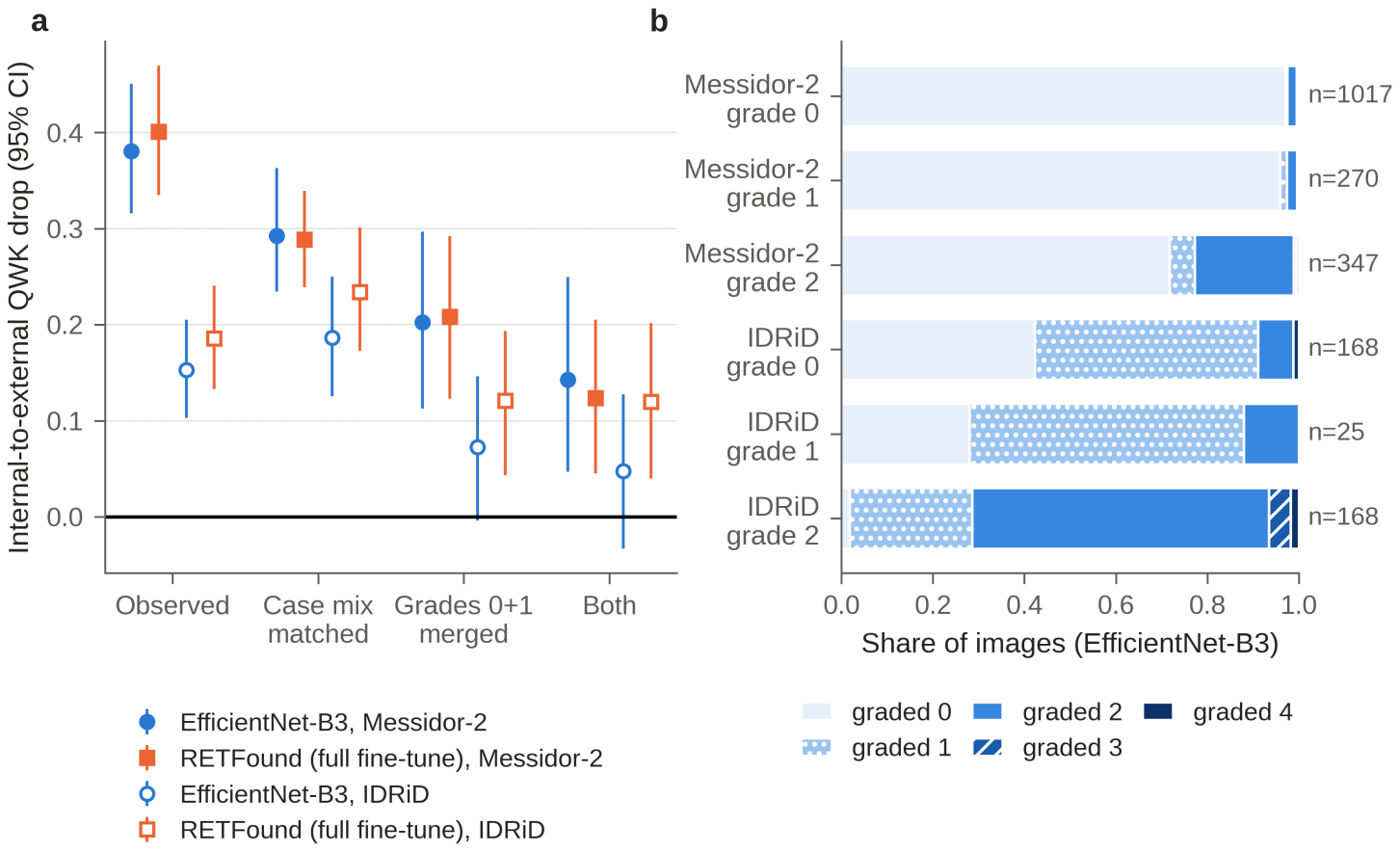

Figure 3

Case mix and the no-DR/mild boundary. **a**Internal-to-external QWK drop for EfficientNet-B3 and fully fine-tuned RETFound as observed, after matching the external case mix to APTOS, after merging grades 0 and 1, and after both corrections (filled markers, Messidor-2; open markers, IDRiD; 95% bootstrap intervals). **b** Grades assigned by EfficientNet-B3 to reference-standard grade 0, 1 and 2 images: mild and moderate DR were graded as no DR on Messidor-2, and no DR was graded as mild on IDRiD

These sensitivities describe a single operating point, each model's highest-ranked grade, and are not suitable for autonomous screening, where missing more than half of referable DR would be unacceptable. Ordinal grading agreement and binary referable-DR discrimination are distinct: the referable-DR AUC of 0.842 for EfficientNet-B3 on Messidor-2 indicates that the models still ranked referable above non-referable images reasonably well, so a lower decision threshold could trade specificity for sensitivity. No such threshold was prespecified.

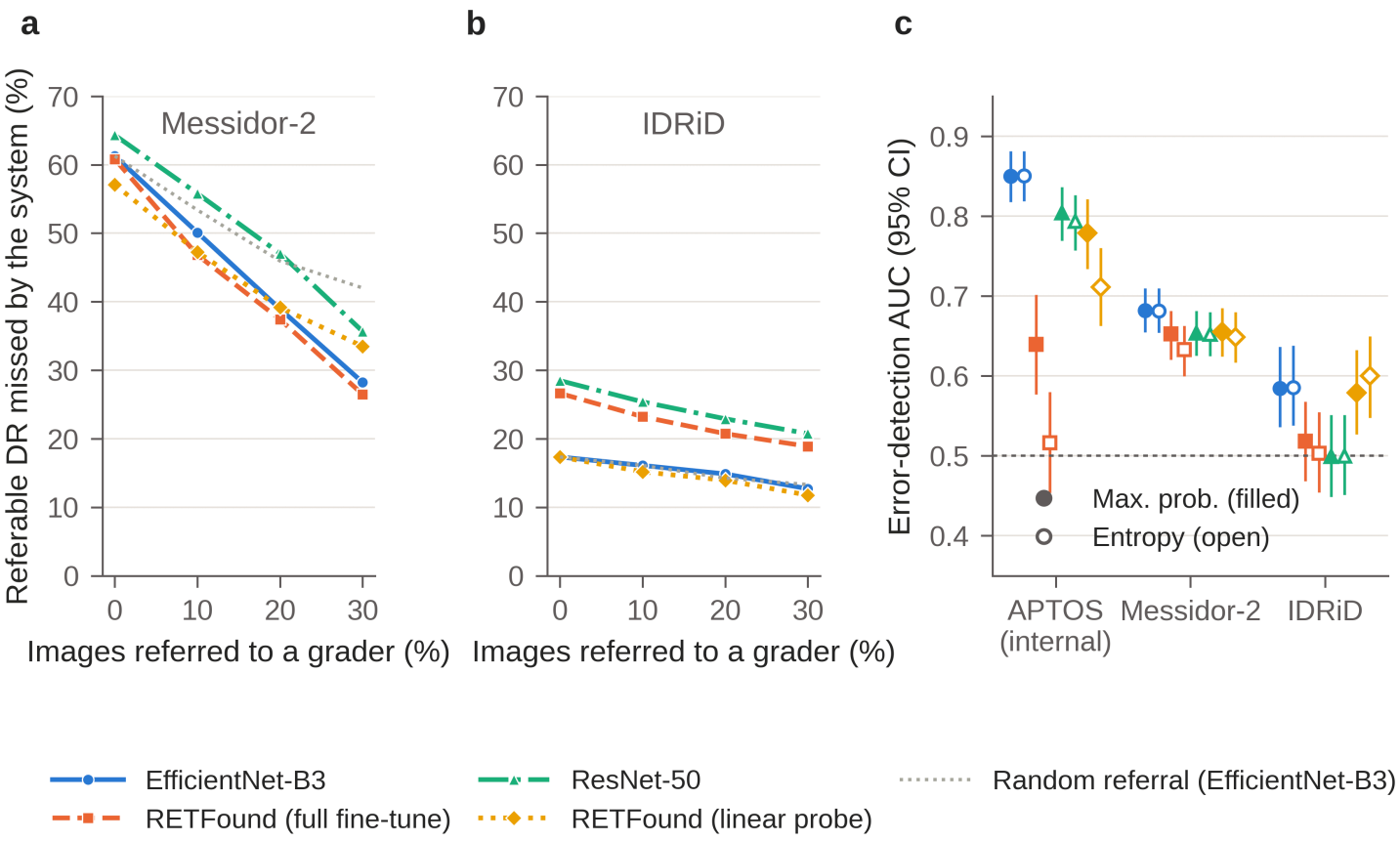

Figure 4

Uncertainty and referral. **a**, **b**Fraction of all referable-DR images missed by the system (retained and graded non-referable) when the least confident images (maximum softmax probability) are referred to a human grader, on Messidor-2 and IDRiD; the dotted line shows a fixed random referral order for EfficientNet-B3. **c** Area under the curve for detecting grading errors with maximum softmax probability (filled) and predictive entropy (open), with 95% bootstrap intervals; the dashed line marks chance

When the least confident 30% of images were referred to a human grader, the fraction of all referable DR missed by the system on Messidor-2 fell from 61% to 28% for EfficientNet-B3 and from 61% to 26% for RETFound, compared with 42% for random referral (Fig. 4a). On IDRiD the reduction was small and no better than random referral (EfficientNet-B3 17% to 13%, random 13%; RETFound 27% to 19%; Fig. 4b). Accuracy on retained Messidor-2 images rose from 63% to 71% for EfficientNet-B3, but QWK on retained images fell from 0.499 to 0.188, because referral preferentially removed diseased images.

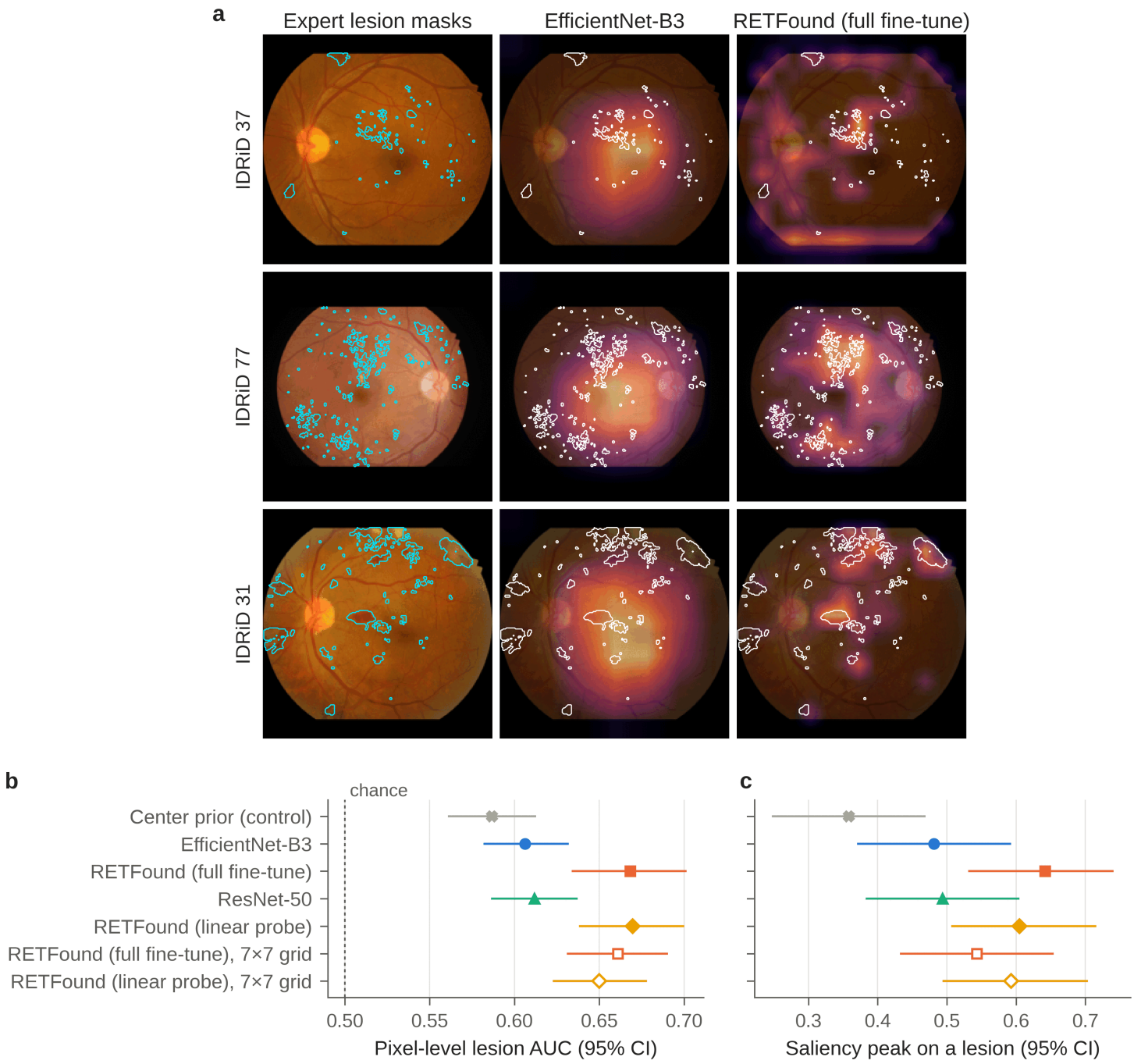

Figure 5

Saliency localization against expert lesion masks (IDRiD Segmentation subset, *n* = 81). **a** Expert lesion outlines and saliency maps for the images at the 25th, 50th and 75th percentiles of lesion burden (selected by rule, not by appearance). **b**Pixel-level AUC for lesion versus non-lesion pixels and **c** fraction of images whose saliency peak lies on a lesion, with 95% bootstrap intervals. Open markers show RETFound maps degraded to the CNN's 7 × 7 resolution; the cross marker is a centered-Gaussian control

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM2.pdf
- ESM1.pdf
- ESM4.zip
- ESM3.zip