diagnostics

MDPI

Article

Label Granularity and Evaluation Design in ODIR-5K: A Leakage-Controlled Benchmark with External Validation for Fundus Disease Classification

Ha Rang Cho [1], Chang-Soo Lee [2] and Woong Sik Kim [3,*]

[1] Department of AI Software Convergence, Konyang University, 158 Gwanjeodong-ro, Seo-gu, Daejeon 35365, Republic of Korea; 26860602@konyang.ac.kr
[2] Korea Research Institute of Bioscience and Biotechnology (KRIBB), 125 Gwahak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea; cslee@kribb.re.kr
[3] Department of Artificial Intelligence, Konyang University, 158 Gwanjeodong-ro, Seo-gu, Daejeon 35365, Republic of Korea
* Correspondence: wskim@konyang.ac.kr

Abstract

Background/Objectives: Accuracies approaching 100% have been reported for fundus disease classification on the ODIR-5K dataset, including 98.2% in our unpublished preliminary pipeline. We quantified how labelling and evaluation shape such figures. **Methods:** Each image was classified as normal or as one of seven disease categories using eye-level labels from ophthalmologists' per-eye keywords (5990 images, 3263 patients). Four convolutional neural networks were fine-tuned, tested externally on RFMiD and locally recalibrated. **Results:** With eye-level labels and patient-level splits, the reference model (EfficientNetB0, 384 pixels) reached a test macro-F1 of 0.659 (95% CI 0.609–0.702) and an accuracy of 70.8% (67.3–74.0%). The preliminary pipeline was trained and evaluated on the same images; its reproduction reached a macro-F1 of 1.000 and 100% accuracy on training images but 0.499 and 53.2% on held-out images. Propagating patient labels to eyes altered 16.4% of labels, and adding fellow eyes to training raised macro-F1 by 0.032 (95% CI −0.004 to 0.068). On four categories shared with RFMiD, macro-F1 and accuracy were 0.815 and 74.9% internally and 0.663 and 75.3% externally. Recalibration raised RFMiD macro-F1 from 0.657 to 0.735 but lowered diabetic retinopathy sensitivity from 0.925 to 0.775. **Conclusions:** Near-perfect accuracy came from evaluation on training images; fellow-eye leakage inflated performance far less.

Keywords: fundus photography; ocular disease classification; data leakage; label noise; ODIR-5K; external validation; deep learning evaluation

Academic Editor: Ayman El-Baz

Received: 17 October 2026
Revised: 19 November 2026
Accepted: 21 November 2026
Published: 24 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Automated analysis of colour fundus photographs can support large-scale screening for sight-threatening disease, and deep convolutional neural networks (CNNs) have matched specialist graders for single conditions such as diabetic retinopathy [1–4]. Screening programmes, however, must recognise several diseases from the same photograph. The Ocular Disease Intelligent Recognition dataset (ODIR-5K) is one of the few public resources that covers multiple diseases in paired left and right fundus images, and it is widely used for this purpose [5–7].

ODIR-5K is annotated at two levels. Each patient carries a set of binary flags for eight categories, while each eye carries free-text diagnostic keywords [5]. Turning the dataset into an image-level, single-label task therefore requires two decisions: which label each image receives, and whether the two eyes of a patient may fall on different sides of a train/test split. Both decisions are rarely reported in detail. On ODIR-5K alone, an accuracy of 99.71% was reported in a study that does not describe how the data were divided into training and test sets or on which set accuracy was measured [8], and a multi-label F1 score of 94.28% in a study that separated labels shared by both eyes into eye-level labels using diagnostic keywords [9]. Accuracies of 95–98% have also been reported on image sets derived in part from ODIR-5K, in one case after a random image-level split of a pooled dataset whose external test images were also drawn partly from ODIR [10]. Where image-level and pseudo-patient-level splits were compared directly on pooled fundus data that included ODIR-5K, macro-F1 fell from 0.931 to 0.623 whereas AUC fell only from 0.976 to 0.934 [11]. In our own unpublished preliminary analysis, a pipeline that assigned each image its patient's highest-priority flag was trained on all 6392 images without a held-out set and reported 98.2% accuracy on those same images. When we later tested the same approach on held-out images, accuracy was far lower, which motivated the present study.

Leakage, in which information unavailable at deployment reaches the model or its evaluation, is a leading cause of irreproducible machine learning results [12,13]. In medical imaging, image-level instead of subject-level splitting inflates accuracy [14], models exploit acquisition shortcuts [15,16], and evaluation design rather than architecture is the main source of over-optimistic claims [17,18]. Labels that do not describe the image add a distinct threat: they corrupt training and make the metric measure agreement with the wrong reference.

This study examines how these factors shape reported performance on ODIR-5K and provides a leakage-controlled reference benchmark. Our contributions are four-fold:

1. We derive eye-level labels from per-eye keywords and show that assigning each image its patient's highest-priority flag yields a label that disagrees with the eye's own keywords for 16.4% of images.
2. We reproduce the original protocol, compare label sources and split units descriptively, and test the effects of label source and of the fellow eye in controlled experiments with the test images held fixed.
3. Under eye-level labels, patient-level splits and validation-only model selection, we benchmark four CNNs and evaluate commonly proposed training choices (class weighting, TensorFloat-32 arithmetic, caching, threshold calibration and input resolution), reporting negative results alongside positive ones.
4. We validate the resulting models on the independent RFMiD dataset [19] under mapping rules fixed before inference, separating the transfer of discrimination from the transfer of operating points, and quantify how much local data is needed to recalibrate them.

2. Materials and Methods

2.1. Dataset and Eye-Level Labelling

In this retrospective study, we used the publicly distributed ODIR-5K training set (6392 colour fundus images from 3358 patients, preprocessed to 512 × 512 pixels) [5], obtained on 16 April 2026 from the Kaggle copy of the ODIR-2019 Grand Challenge data (annotation file full_df.csv and folder preprocessed_images; see the Data Availability Statement). Each patient has eight binary flags (normal, diabetic retinopathy, glaucoma, cataract, age-related macular degeneration [AMD], hypertensive retinopathy, myopia,

other), and each eye has comma-separated diagnostic keywords; patient age and sex are also provided. The images were acquired by the provider with different fundus cameras at multiple ophthalmic centres in China [5]; no further acquisition parameters are available.

The per-eye keywords were assigned by the provider's annotation team: three ophthalmologists annotated each image independently, disagreements were arbitrated by three senior ophthalmologists, and final labels required agreement of at least two experts [5]. Eye-level labels were derived from each eye's own keywords with a fixed dictionary (Supplementary Table S1). Keywords describing image quality (e.g., lens dust, low image quality) were ignored. Keywords marked "suspected" or "suspicious" led to exclusion. Laser photocoagulation scars were ignored when another diagnosis was present and were otherwise mapped to other. Images whose keywords mapped to more than one category were excluded from the single-label task. The final dataset comprised 5990 images from 3263 patients (Table 1). To check the dictionary against an expert reference independent of the authors, we compared it with the provider's patient-level category flags: for each keyword that was the only abnormal keyword in both eyes of a patient, the category flagged by the provider was taken as the provider's mapping of that keyword and was compared with ours.

Table 1. Eye-level class distribution after exclusions and patient-level split. Excluded: 356 multi-disease and 46 uncertain-diagnosis images.

Class	Development (5 Folds), Images	Test, Images	Total, Images
Normal	2463	410	2873
Diabetic retinopathy	1220	204	1424
Glaucoma	174	30	204
Cataract	238	39	277
AMD	204	34	238
Hypertensive retinopathy	89	15	104
Myopia	200	33	233
Other	546	91	637
Total (patients)	5134 (2797)	856 (466)	5990 (3263)

For comparison, we reconstructed the propagated patient-level labels used in our preliminary analysis: each image received the first positive patient flag in the order normal > diabetic retinopathy > glaucoma > cataract > AMD > hypertension > myopia > other. This rule reproduced the class folders of the preliminary analysis for all 6392 images. The preliminary model (EfficientNetB0 with ImageNet weights, 224 pixels after an 85% central crop, per-image standardisation, class-weighted loss, batch size 64, 35 epochs, no augmentation) was trained on all 6392 images without a validation or test split, and its 98.2% was the training accuracy reported at the end of the final epoch; it was therefore trained and evaluated on the same images. Its training script and log are provided in Supplementary Data S6.

2.2. Patient-Level Splitting

Images were split with stratified group k-fold sampling (scikit-learn version 1.6.1), using the patient as the group and the eye-level class for stratification. One seventh of patients formed a hold-out test set (856 images, 466 patients); the remaining 5134 images (2797 patients) were divided into five cross-validation folds. No patient appears in more than one partition (Figure 1a). The test set was not used for any training, tuning or selection decision.

Duplicate images were screened in all 6392 ODIR-5K and 3200 RFMiD images. Exact duplicates were sought by file checksum and near-duplicates by a 64-bit perceptual hash of each image and its mirror image. Because the pre-specified hash threshold (Hamming distance ≤ 6) proved non-discriminative for fundus photographs (38,660 candidate pairs), review was restricted post hoc to ODIR patient pairs in which both eyes matched (Hamming distance ≤ 4, or ≤6 when the pair crossed the development–test boundary; 34 pairs) and to ODIR–RFMiD pairs with a distance ≤ 2 (36 pairs). Each candidate was classified by the correlation of high-pass-filtered green-channel images within the fundus mask (a vessel-pattern fingerprint, maximised over shifts of up to 10 pixels) and by visual review; confirmed duplicates crossing the development–test boundary were removed from the test set in a sensitivity analysis.

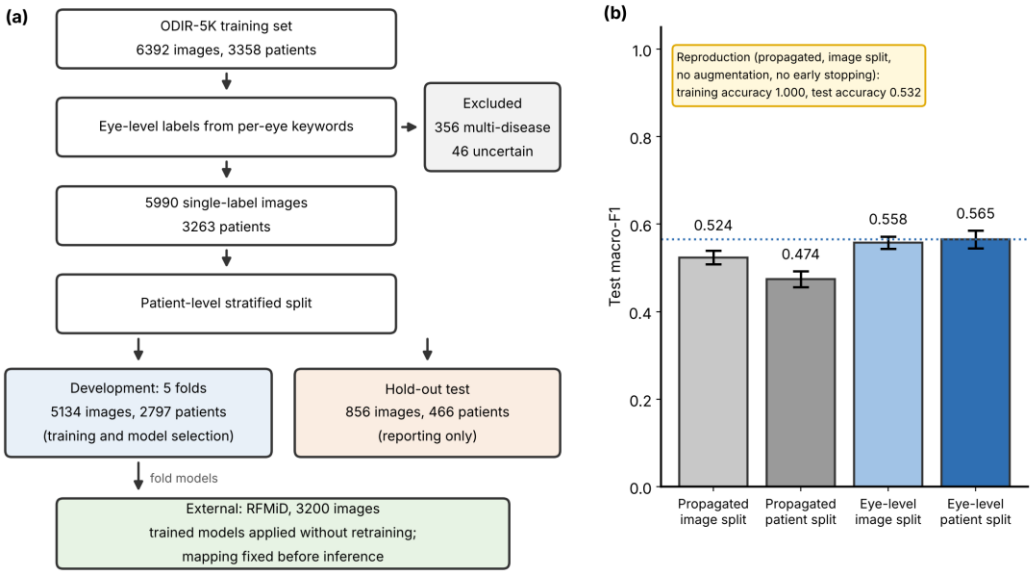


Figure 1. Study design. (**a**) Data flow: 6392 images from 3358 patients → eye-level labelling from per-eye keywords → exclusion of 356 multi-disease and 46 uncertain images → 5990 images (3263 patients) → patient-level split into a hold-out test set (856 images, 466 patients) and five development folds (5134 images, 2797 patients); external validation on RFMiD (3200 images) without updating network weights. (**b**) Descriptive comparison of label source (propagated patient-level versus eye-level) and split unit (image versus patient), plus a reproduction of the preliminary protocol; values are test macro-F1 (mean of five folds; reproduction, single run); error bars, ±1 SD across folds; blue dotted line, eye-level labels with patient-level split (this study, 0.565).

2.3. Preprocessing and Models

Images were resized with bicubic interpolation to 224 or 384 pixels and passed through each backbone's own ImageNet preprocessing. Training augmentation consisted of horizontal flips, rotation (±18°), zoom (±10%), brightness (±10%) and contrast (±10%), applied per image after caching.

We compared EfficientNetB0 [20], DenseNet121 [21], ResNet50V2 [22] and MobileNetV2 [23], all initialised with ImageNet weights [24]. EfficientNetV2-S, a more recent and larger architecture, was added as a sensitivity analysis. Each backbone was followed by the same head: global average pooling, dropout (0.3), a 256-unit ReLU layer, dropout (0.3) and an eight-way softmax.

2.4. Training Protocol

Training had two stages. The head was first trained for three epochs with the backbone frozen (Adam [25], learning rate 1×10^{-3}). The whole network was then fine-

tuned at 1×10^{-4} for up to 60 epochs, with batch normalisation kept in inference mode, batch size 32 and categorical cross-entropy loss. The learning rate was reduced by a factor of 0.3 after four epochs without improvement in validation macro-F1, and training stopped after ten. The checkpoint with the highest validation macro-F1 was retained. Class weights [26,27], when used, were $w_i = N/(k \cdot n_i)$, computed on the training folds only, where N is the number of training images, $k = 8$ and n_i is the count of class i.

2.5. Experiments

Experiments compared the four backbones at 224 pixels; evaluated class weighting, TensorFloat-32 (TF32) arithmetic [28], caching and, for EfficientNetB0 and DenseNet121, a 384-pixel input; tested class-wise threshold calibration fitted on out-of-fold predictions; and compared label sources (propagated, eye-level) and split units (image, patient) (Figure 1b), including a reproduction of the preliminary protocol without augmentation or early stopping, to which an image-level split was added so that training-set and held-out performance could be compared. EfficientNetB0 at 384 pixels, the reference configuration chosen before the sensitivity analyses, was repeated with three seeds. EfficientNetV2-S was trained at 384 pixels with batch size 16 because batch 32 exceeded GPU memory; ConvNeXt-Tiny was attempted but could not be compiled in this environment (Supplementary Text S1). Grad-CAM [29] maps from the final model were used for qualitative illustration only. The full experimental grid is given in Supplementary Text S3, and per-run metrics in Supplementary Data S1. Two controlled experiments were added, both with EfficientNetB0 at 224 pixels, seed 0 and the settings of the reference runs. First, to compare label sources on identical data, models were retrained on the same 5990 images and folds with propagated labels (model selection on propagated validation labels) and scored against eye-level labels on the same 856 test images. Second, to isolate fellow-eye leakage (seeds 0–2, with the reference runs also repeated with seeds 1 and 2), one randomly chosen eye of each of the 390 test patients with two single-label eyes was fixed as an evaluation eye; its fellow eye was added to the training partition of every fold model, and the same number of training images of the same classes was removed, so that training-set size, class counts and validation folds were unchanged. Both experiments were compared with the reference runs on the same test images with paired bootstrap tests; for the fellow-eye experiment, the 15 models of each arm (three seeds × five folds) were averaged into one ensemble per arm. Training at the native 512-pixel resolution was attempted but exceeded the 16 GB GPU memory at the protocol's batch size (Limitations Section). The propagated-label models were also applied to RFMiD without retraining and compared with the eye-level models on the four shared categories.

2.6. External Validation and Excluded Multi-Disease Images

External validation used the Retinal Fundus Multi-disease Image Dataset (RFMiD; 3200 images from three fundus cameras, annotated by consensus of two retinal specialists) [19], obtained from a Kaggle copy of the original release (see the Data Availability Statement). The primary external validation used frozen models: RFMiD labels were not used for training, fine-tuning, model selection or threshold setting. A separate local recalibration analysis, described below, then fitted class-wise offsets on the official RFMiD training split. Throughout, "no retraining" means that network weights were not updated. Mapping rules were recorded with a time stamp and script hash before inference (Supplementary Text S2): Disease_Risk = 0 with no disease label → normal (images with Disease_Risk = 1 but no disease label were excluded as ambiguous); DR → diabetic retinopathy; ARMD → AMD; MYA → myopia; optic disc cupping → glaucoma (approximate; secondary analyses only). RFMiD has no cataract label, and its HR label denotes haemorrhagic rather than hypertensive retinopathy; neither class was mapped.

Images were cropped to the fundus and resized as for ODIR-5K. The public RFMiD release provides no patient age or sex.

Three analysis populations were defined in advance (Supplementary Table S8). Per-class one-versus-rest AUC used all 3200 images, including those with several labels. Single-label analyses used 2451 images: normal images (Disease_Risk = 0 with no disease label) and images with exactly one disease annotation (a single unmapped disease → other); images with two or more disease labels, or with Disease_Risk = 1 but no disease label, were excluded. The primary comparison used the 1214 images that belonged to the four diagnostic categories shared by both datasets (669 normal, 401 diabetic retinopathy, 69 AMD, 75 myopia). The complete column-by-column label mapping is given in Supplementary Table S9. Predictions kept all eight ODIR classes, so predicting an unevaluable class counted as an error. In the four-category analysis, macro-F1 and balanced accuracy were averaged over the four evaluated categories, so an image predicted as any other ODIR class counted as a false negative for its true category; one-versus-rest AUC used the eight-class softmax probability of each evaluated category. For the primary comparison, the same metrics were computed on the ODIR test set restricted to those four classes (681 images), with patient-level bootstrap CIs for ODIR and image-level bootstrap CIs for RFMiD, which lacks patient identifiers. As a sensitivity analysis for the single-label design, the 356 excluded multi-disease ODIR images were scored with the EfficientNetB0 (384 pixels) five-fold ensemble by whether the top-ranked prediction matched any of their diagnoses; chance agreement was the mean over images of the number of diagnoses divided by eight. Results were also stratified by the location of the patient's fellow eye (development set, hold-out test set, or no single-label fellow eye).

Local recalibration was a separate analysis that used RFMiD labels and followed a second protocol, also fixed before computation (Supplementary Text S2). Class-wise additive offsets b_c were added to the log-probabilities, and the final prediction was the class with the largest adjusted score, $\hat{y} = \text{argmax}_c (\log p_c + b_c)$, with offsets fitted for all eight ODIR classes so that predictions outside the four shared categories remained possible. The offsets were initialised at zero and fitted by coordinate ascent over a grid from −4 to 4 (step 0.1) for up to three passes over the classes with early stopping when no offset changed; ties were broken in favour of the smaller $|b_c|$, and the offsets were then centred to zero mean, which does not affect the argmax. The objective was macro-F1 over the four shared categories of the official RFMiD training split (721 images); network weights were not changed. Evaluation used the official test split (249 images), with the validation split (244 images) as a second independent check. After recalibration, macro AUC was computed from the offset-adjusted, renormalised probabilities, softmax(log p + b); because renormalisation after class-wise offsets can change the relative ordering of images within each one-versus-rest comparison, AUC can change slightly. We compared no recalibration, restriction of predictions to the four evaluated categories (probabilities of the other four ODIR classes set to zero and renormalised; no local data needed), offset fitting, and restriction combined with offset fitting. As a sensitivity analysis, offsets were also fitted to maximise macro-F1 subject to a minimum diabetic retinopathy sensitivity of 0.90 or 0.95 on the training split, rejecting candidate offsets that violated the constraint. To estimate how much local data was needed, offsets were refitted on stratified random subsets of 25 to 400 training images (20 repeats each). Confidence intervals used 2000 image-level bootstrap resamples of each evaluation split; all four categories were present in every resample of the test and validation splits. Paired differences between no recalibration and offset fitting in macro-F1, balanced accuracy and macro AUC were estimated on the same resamples.

https://doi.org/10.3390/diagnostics16233957

2.7. Evaluation and Statistics

The primary metric was macro-F1; secondary metrics were balanced accuracy (mean per-class sensitivity), accuracy and macro one-versus-rest AUC. For each setting, we report the mean ± standard deviation over the five fold models and the performance of their probability-averaged ensemble; "ensemble" denotes the average of the five-fold-model probabilities of one seed, and values over individual fold models are labelled as such. Class-specific sensitivity, precision, F1 and one-versus-rest AUC were reported with patient-level bootstrap confidence intervals. Performance was also summarised by sex and by age group (<50, 50–64 and ≥65 years); within each subgroup, metrics were averaged over the classes present, and bootstrap resamples were restricted to those containing the subgroup's full class set. Confidence intervals (95%) were obtained by patient-level bootstrap with 2000 resamples [30]. Settings sharing a test set were compared with paired patient-level bootstrap tests (two-sided), with Holm correction for the three model comparisons at 224 pixels (EfficientNetB0 versus each other backbone) [31]; all other tests were not corrected for multiplicity and are exploratory. Two-sided bootstrap p-values were computed as twice the smaller of the proportions of resampled differences ≤ 0 and ≥0 (zero differences counted in both tails, without a +1 correction) and capped at 1. When no resampled difference fell on the opposite side of zero, this estimate is zero (stored as 1/2000 in the analysis output); such values are reported as $p < 0.001$, the smallest non-zero value attainable with 2000 resamples. For ODIR-5K, macro-F1 was averaged over the fixed set of eight classes, and resamples in which any one-versus-rest AUC was undefined were excluded from the AUC estimates; for RFMiD, metrics were averaged over the four evaluated categories, all of which were present in every resample of the four-category analyses. All model and setting decisions were based on out-of-fold validation predictions; test results are reported for confirmation only. Epoch time is the median over fine-tuning epochs excluding each stage's first (compilation) epoch; convergence is the first epoch reaching 95% of maximum validation macro-F1.

2.8. Implementation

Experiments ran on one NVIDIA GeForce RTX 5080 (16 GB, compute capability 12.0; NVIDIA Corporation, Santa Clara, CA, USA) under Windows 11 (Microsoft Corporation, Redmond, WA, USA) with Python 3.9 (Python Software Foundation, Wilmington, DE, USA), TensorFlow 2.10.1 (Google LLC, Mountain View, CA, USA), CUDA 11.2 and cuDNN 8.1 (NVIDIA Corporation). This TensorFlow build contains native kernels up to compute capability 7.5 (sm_75) and PTX for compute_80, so on this GPU, TensorFlow's kernels were compiled from PTX by the driver at run time, a configuration outside the GPU generations supported by CUDA 11.2. Numerical diagnostics are described in Supplementary Text S1. Every run stores its configuration (including seed, software versions and the runtime TF32 state), training history and predictions in the repository; runs that ended with non-finite losses were recorded as failed and were not repeated. The additional revision experiments ran in the same environment, with a separate run directory, and all their configurations, hashes and deviations from the revision protocol are archived; their per-run metrics are given in Supplementary Data S5.

2.9. Use of Generative AI

The authors designed this study, directed every step and made all methodological and editorial decisions. During the preparation of this study, the authors used Claude (Anthropic, San Francisco, CA, USA; Claude Code, Claude Fable 5.2, 2026) as an assistant under their direction to assist with code development for the training and evaluation pipeline, to review the experimental design, and to draft the manuscript text from the authors' outlines and notes; during the preparation of this version, to retrieve and check

the cited literature, to draft the revised abstract and inserted passages from the authors' instructions, to convert the manuscript to the journal format, to write analysis scripts and to assist the classification of duplicate-image candidates, which the authors checked; and for language editing. The authors executed every experiment, checked every reported value against the result files, and revised and take full responsibility for all code and text.

3. Results

3.1. Eye-Level Labels

Of the 6392 images, 356 had keywords spanning more than one category and 46 carried uncertain diagnoses; 5990 images remained (Table 1). Patients were aged 57.8 ± 11.7 years (median 59, interquartile range 51–66; range 1–91 as recorded in the annotation file, with 18 patients younger than 18 years), and 46.5% were female; the development and test sets were similar (Supplementary Table S15). The dictionary agreed with the provider's patient-level category for all 22 keywords that occurred as the only abnormal keyword of a patient (1465 patients; $\kappa = 1.0$), including judgement-dependent keywords such as drusen and epiretinal membrane, which the provider also assigned to other; images whose keywords were all covered by this check made up 99.1% of the dataset (Supplementary Table S18). The propagated patient-level label differed from the eye-level label for 16.4% of images (981 of 5990). Of these, 772 (78.7%) were eyes whose own keyword was "normal fundus" but whose fellow eye was diseased, so this disagreement does not depend on the dictionary; alternative mappings of judgement-dependent keywords gave 15.9–18.4%. Our eye-level labels agreed with the distributor's per-image label, which reflects only the first keyword, for 99.9% of the 5990 single-label images (93.7% of all 6392 images when excluded images are counted as disagreements; Supplementary Table S2).

No exact duplicate files were found. Of the 34 bilateral patient-pair candidates, five were confirmed as the same two eyes registered under different patient identifiers (vessel-pattern correlation 0.64–1.00, against at most 0.36 for the other 29 pairs): two pairs of re-compressed copies of the same photographs (patients 2487/3185 and 352/973) and three pairs of near-identical photographs (271/282, 32/765 and 398/668). Two pairs crossed the development–test boundary, affecting four test images; removing them changed the test macro-F1 of the EfficientNetB0 (384 pixels) ensemble from 0.659 to 0.660 and left the macro AUC unchanged (0.913). One within-development pair carried different categories for the same eyes (hypertensive retinopathy versus diabetic retinopathy). None of the 36 ODIR–RFMiD candidates was a duplicate (vessel-pattern correlation ≤ 0.27; Supplementary Table S23).

3.2. Label Source, Split Unit and Fellow-Eye Leakage

Table 2 compares the conditions descriptively. Reproducing the preliminary protocol drove training-set accuracy and macro-F1 to 1.000, whereas the same model reached an accuracy of 0.532 and a macro-F1 of 0.499 on held-out images; on all 6392 images, 68.6% of which were its training images, accuracy was 0.851 (Supplementary Data S6). With this held-out accuracy, 98.2% would require about 96% of the evaluated images to be training images, consistent with the preliminary model having been trained and evaluated on all of them (Section 2.1). Because the reproduction used other training settings and a split, it illustrates rather than replicates the preliminary figure. Its held-out results use propagated labels and an image-level test set and are therefore not directly comparable with results under eye-level labels on the patient-level test set (Table 3).

Table 2. Descriptive comparison of label source and split unit (EfficientNetB0, 224 pixels). Values are mean ± SD over five folds; the reproduction is a single run. Test sets differ between label sources, so rows are compared descriptively.

Labels	Split	Fellow Eye in Development Set	Training-Set Accuracy	Test Accuracy	Test Macro-F1	Test Macro AUC	Macro-F1 vs. Eye-Level Truth
Propagated	Image, no augmentation or early stopping (reproduction)	81.4%	1.000	0.532	0.499	0.811	0.513
Propagated	Image	81.4%	0.824 ± 0.087	0.521 ± 0.009	0.524 ± 0.015	0.830 ± 0.010	0.540 ± 0.022
Propagated	Patient	0%	0.747 ± 0.106	0.468 ± 0.019	0.474 ± 0.018	0.806 ± 0.009	0.523 ± 0.023
Eye-level	Image	76.5%	0.905 ± 0.022	0.615 ± 0.009	0.558 ± 0.014	0.858 ± 0.008	0.558 ± 0.014
Eye-level	Patient (this study)	0%	–	0.608 ± 0.033	0.565 ± 0.020	0.878 ± 0.004	0.565 ± 0.020

Table 3. Model comparison on the hold-out test set. Fold columns: Mean ± SD of five fold models. Ensemble: Average of fold-model probabilities, with patient-level bootstrap 95% CI. EfficientNetB0 rows use seed 0; the 15 individual EfficientNetB0 models at 384 pixels (five folds × three seeds) gave macro-F1 0.631 ± 0.019, balanced accuracy 0.625 ± 0.018 and macro AUC 0.899 ± 0.007, and the three seed-wise five-fold ensembles gave macro-F1 0.661 ± 0.002 and macro AUC 0.913 ± 0.004. † Sensitivity analysis, batch size 16.

Model	Input (px)	Params (M)	Macro-F1 (Folds)	Balanced Accuracy (Folds)	Macro AUC (Folds)	Ensemble Macro-F1 [95% CI]	Ensemble Macro AUC [95% CI]	Ensemble Accuracy [95% CI]	Epoch Time (s)
EfficientNetB0	224	4.38	0.565 ± 0.020	0.567 ± 0.032	0.878 ± 0.004	0.581 [0.533–0.624]	0.897 [0.878–0.915]	0.638 [0.602–0.672]	8.7
DenseNet121	224	7.30	0.547 ± 0.019	0.556 ± 0.018	0.850 ± 0.008	0.586 [0.538–0.629]	0.874 [0.849–0.895]	0.583 [0.543–0.621]	14.3
MobileNetV2	224	2.59	0.536 ± 0.023	0.565 ± 0.014	0.846 ± 0.006	0.563 [0.521–0.602]	0.872 [0.850–0.892]	0.576 [0.538–0.612]	5.5
ResNet50V2	224	24.09	0.513 ± 0.024	0.502 ± 0.023	0.831 ± 0.015	0.547 [0.503–0.584]	0.867 [0.843–0.886]	0.562 [0.524–0.596]	10.7
EfficientNetB0	384	4.38	0.636 ± 0.021	0.627 ± 0.014	0.898 ± 0.010	0.659 [0.609–0.702]	0.913 [0.895–0.931]	0.708 [0.673–0.740]	26.6
DenseNet121	384	7.30	0.591 ± 0.016	0.605 ± 0.033	0.881 ± 0.008	0.627 [0.572–0.673]	0.904 [0.882–0.922]	0.667 [0.630–0.700]	35.1
EfficientNetV2-S †	384	20.66	0.638 ± 0.008	0.632 ± 0.017	0.910 ± 0.011	0.683 [0.637–0.724]	0.930 [0.912–0.944]	0.759 [0.727–0.789]	40.2

Under propagated labels, an image-level split placed the fellow eyes of 81.4% of the test images in the development set, always with the same label. Held-out macro-F1 was 0.524 with the image-level split and 0.474 with the patient-level split; because the two splits produce different test sets, this difference cannot be attributed to fellow-eye leakage alone. Under eye-level labels, macro-F1 was similar for the two splits (0.558 versus 0.565), and fellow eyes shared a label for only 54.6% of test images. Models trained on propagated labels and re-scored against eye-level diagnoses reached a macro-F1 of 0.523, compared with 0.565 for models trained on eye-level labels, although their test sets also differed.

In the controlled comparisons (Supplementary Table S21), models trained on propagated labels and scored on the same 856 test images against eye-level labels reached a macro-F1 of 0.558 versus 0.581 for eye-level training (difference 0.022, 95% CI −0.009 to 0.054; $p = 0.15$) and a macro AUC of 0.877 versus 0.897 (difference 0.020, 0.007 to 0.034; $p = 0.008$). With the 390 evaluation eyes fixed and training-set size matched, adding the test patients' fellow eyes to training raised macro-F1 in all three seeds (by 0.047, 0.027 and 0.022); for the 15-model ensembles, macro-F1 was 0.636 versus 0.604 (difference 0.032,

−0.004 to 0.068; $p = 0.08$) and accuracy 0.697 versus 0.667 (difference 0.031, 0.003 to 0.059), whereas macro AUC changed little (0.912 versus 0.900; difference 0.012, −0.002 to 0.027). Training-set size and class counts were matched, but the number of distinct patients was not: the fellow-eye training sets contained 333–341 more patients per fold (2571–2578 versus 2237–2238), so part of this gain may reflect more varied training data. The accuracy gain was larger for evaluation eyes whose fellow eye had the same diagnosis (0.039, 95% CI 0.004 to 0.074; 282 eyes) than for those whose fellow eye had a different diagnosis (0.009, −0.037 to 0.056; 108 eyes); macro-F1 within these subgroups was unstable because several classes had only a few images (differences 0.032 and 0.070, both intervals including zero; Supplementary Data S6). Differences were computed from unrounded values and may therefore differ by 0.001 from the differences between the rounded values given here.

3.3. Model Comparison

At 224 pixels, EfficientNetB0 achieved the highest out-of-fold validation performance and the highest fold-mean test macro-F1 and macro AUC (Table 3), although the DenseNet121 ensemble had a slightly higher test macro-F1 (0.586 versus 0.581). Its macro AUC exceeded that of DenseNet121, MobileNetV2 and ResNet50V2 by 0.024 to 0.030 (paired bootstrap $p = 0.036$, 0.006 and 0.003; all significant after Holm correction; Supplementary Table S5; differences were computed from unrounded values and may differ by 0.001 from differences between the rounded entries in Table 3). Differences in macro-F1 and balanced accuracy were not significant. EfficientNetB0 also converged earliest (95% of best validation macro-F1 at 18.0 ± 4.1 epochs versus 22.8 to 28.4).

3.4. Input Resolution

Increasing the input to 384 pixels improved out-of-fold validation macro-F1 by 0.048 for EfficientNetB0 and 0.051 for DenseNet121, and macro AUC by 0.046 and 0.038. On the test set, the EfficientNetB0 ensemble improved in macro-F1 from 0.581 to 0.659 (difference 0.078, 95% CI 0.041–0.121, $p < 0.001$) and in macro AUC from 0.897 to 0.913 ($p = 0.008$). DenseNet121 improved significantly in macro AUC (0.874 to 0.904, $p < 0.001$) but not in macro-F1 ($p = 0.118$). At 384 pixels, the two architectures did not differ significantly in macro-F1, balanced accuracy or macro AUC ($p = 0.15$ to 0.87); EfficientNetB0 used 40% fewer parameters and 24% less time per epoch in our environment (26.6 versus 35.1 s). Across the 15 individual models (five folds × three seeds), EfficientNetB0 at 384 pixels reached a test macro-F1 of 0.631 ± 0.019 and macro AUC of 0.899 ± 0.007, whereas the five-fold ensembles of the three seeds reached macro-F1 of 0.659, 0.662 and 0.661 (0.661 ± 0.002) and macro AUC of 0.913 ± 0.004 (Supplementary Table S20). The reported 0.659 is therefore an ensemble value typical of all three seeds, and its difference from 0.631 reflects ensembling rather than seed selection. EfficientNetV2-S (20.7 M parameters, 40.2 s per epoch) did not differ significantly from EfficientNetB0 in macro-F1 (0.683 versus 0.659, $p = 0.20$) or macro AUC (0.930 versus 0.913, $p = 0.061$).

Gains were concentrated in classes defined by small lesions (Table 4). Diabetic retinopathy sensitivity rose from 0.48 to 0.66 and AMD from 0.59 to 0.77, whereas cataract and myopia, recognised from global appearance, did not improve. Hypertensive retinopathy remained largely undetected (2 of 15 test images). Class-specific intervals were wide for the minority classes, with sensitivity 0.133 (95% CI 0.000–0.300) for hypertensive retinopathy and 0.567 (0.385–0.762) for glaucoma (Table 4; Supplementary Table S16).

Table 4. Per-class test performance of the EfficientNetB0 ensemble at 384 versus 224 pixels. Sensitivity 95% CI: Patient-level bootstrap (2000 resamples) for the 384-pixel ensemble; precision, F1 and AUC intervals are given in Supplementary Table S16.

Class	Test Images	Correct, 384/224	Sensitivity, 384/224	Sensitivity 95% CI, 384	F1, 384/224	AUC, 384/224
Normal	410	321/311	0.783/0.759	0.737–0.824	0.753/0.711	0.836/0.800
Diabetic retinopathy	204	135/98	0.662/0.480	0.586–0.735	0.673/0.531	0.881/0.821
Glaucoma	30	17/14	0.567/0.467	0.385–0.762	0.596/0.538	0.933/0.928
Cataract	39	34/36	0.872/0.923	0.743–0.972	0.840/0.878	0.996/0.996
AMD	34	26/20	0.765/0.588	0.600–0.926	0.825/0.563	0.963/0.944
Hypertensive retinopathy	15	2/1	0.133/0.067	0.000–0.300	0.200/0.105	0.861/0.852
Myopia	33	28/29	0.848/0.879	0.682–0.972	0.889/0.906	0.995/0.998
Other	91	43/37	0.473/0.407	0.358–0.580	0.494/0.411	0.841/0.838

3.5. Training Settings

No training setting improved macro-F1 (Table 5); with TF32 disabled, four of the five class-weighted runs failed, so that configuration has no five-fold estimate. Class weighting traded accuracy (−0.04) for balanced accuracy (+0.02) and delayed the best epoch from about 16 to 29. In this environment, TF32 changed neither epoch time nor accuracy when training succeeded, and caching shortened epochs by 7.5%. Threshold calibration did not consistently improve nested out-of-fold macro-F1. At 224 pixels in the reference configuration (class weighting and TF32 enabled), it reduced test balanced accuracy for all four backbones (unadjusted paired-bootstrap $p \leq 0.013$; exploratory); calibration was therefore not adopted (Supplementary Table S4).

Table 5. Training settings for EfficientNetB0 at 224 pixels (test, mean ± SD over five folds). Caching (fold 0, three seeds): 8.73 s per epoch with caching versus 9.44 s without.

Class Weighting	TF32	Completed Runs	Macro-F1	Balanced Accuracy	Accuracy	Macro AUC	Best Epoch	Epoch Time (s)
Balanced	On	5/5	0.565 ± 0.020	0.567 ± 0.032	0.608 ± 0.033	0.878 ± 0.004	28.6	8.73
None	On	5/5	0.572 ± 0.020	0.550 ± 0.017	0.644 ± 0.023	0.885 ± 0.005	16.4	8.88
None	Off	5/5	0.563 ± 0.017	0.544 ± 0.023	0.644 ± 0.012	0.885 ± 0.006	15.4	8.84
Balanced	Off	1/5 (4 non-finite)	—	—	—	—	—	—

Completed runs = runs that finished without non-finite losses. For the class-weighted TF32-off configuration, only one of five runs completed, so no five-fold summary is given (—); a single fold would not be comparable with the five-fold means of the other rows. The completed run (fold 2) reached a test macro-F1 of 0.600, balanced accuracy of 0.591, accuracy of 0.640 and macro AUC of 0.873 (best epoch 27; 8.65 s per epoch); the class-weighted, TF32-enabled run of the same fold reached 0.597, 0.593, 0.634 and 0.878 (best epoch 32; 8.73 s). The four failed runs are listed in Supplementary Table S3.

3.6. Numerical Failures

With TF32 disabled, some runs ended with non-finite losses (4 of 5 class-weighted EfficientNetB0 runs and 4 of 6 MobileNetV2 runs, against none of 17 comparable runs with TF32 enabled); the diagnostics are given in Supplementary Text S1 and Table S3.

3.7. Qualitative Explanations

The fold-0 model classified 14 of 24 sampled images correctly, none of them glaucoma or hypertension. Activations overlapped drusen and hard exudates in some correct cases but fell outside the field of view in one glaucoma image (Figure 2).

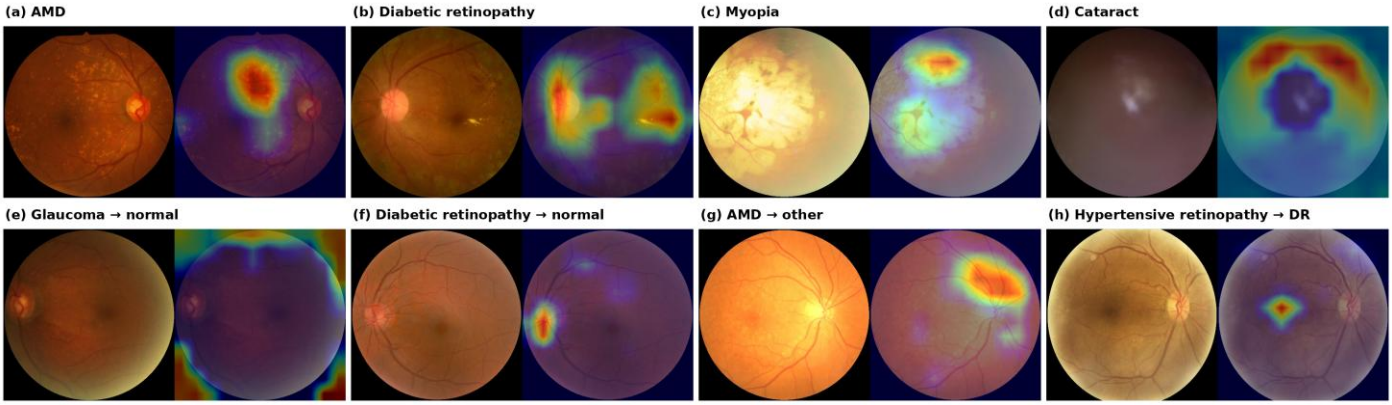


Figure 2. Grad-CAM maps of the final model (EfficientNetB0, 384 pixels, fold 0) for test images selected from a seeded random sample of 24 disease-class images (four per class for diabetic retinopathy, glaucoma, cataract, AMD, hypertensive retinopathy and myopia). Each panel shows the input (**left**) and the overlay (**right**). Top row, correctly classified: (**a**) AMD, (**b**) diabetic retinopathy, (**c**) myopia, (**d**) cataract. Bottom row, misclassified (true → predicted): (**e**) Glaucoma → normal, (**f**) diabetic retinopathy → normal, (**g**) AMD → other, (**h**) hypertensive retinopathy → diabetic retinopathy. Maps are computed for the true class; for misclassified images, the true-class probability was ≤0.07, so these maps show weak evidence. The 12 × 12 feature resolution permits only coarse localisation, and in (**e**), activation lies outside the field of view. No expert annotation was performed; the maps are illustrative only.

3.8. External Validation

On the four shared classes (1214 single-label RFMiD images), the EfficientNetB0 (384 pixels) ensemble reached a macro AUC of 0.939 (95% CI 0.925–0.953) on RFMiD, compared with 0.918 (0.896–0.937) on the ODIR test set, with overlapping intervals. In contrast, macro-F1 fell from 0.815 (0.769–0.852) to 0.663 (0.617–0.707) and balanced accuracy from 0.764 (0.705–0.820) to 0.612 (0.571–0.654), with non-overlapping intervals (Table 6). Accuracy, by contrast, was similar on the two datasets (0.749 on the 681 shared-category ODIR test images and 0.753, 95% CI 0.727–0.778, on RFMiD): normal and diabetic retinopathy images, whose F1 changed little (normal 0.801 and 0.800; diabetic retinopathy 0.701 and 0.771), made up 90% and 88% of the two sets, whereas macro-F1 weights the four categories equally and fell mainly through AMD (0.852 to 0.617) and myopia (0.903 to 0.465) (Supplementary Data S6). Across all 3200 RFMiD images, including multi-label images, one-versus-rest AUC was 0.917 (95% CI 0.903–0.930) for diabetic retinopathy, 0.863 (0.828–0.894) for AMD, 0.857 (0.823–0.888) for myopia, 0.796 (0.779–0.811) for normal and 0.748 (0.720–0.774) for optic disc cupping (Table 6b; Supplementary Table S6). Adding the approximate glaucoma class and then the heterogeneous other class (2451 single-label images) lowered external macro-F1 to 0.547 (0.507–0.584) and 0.393 (0.360–0.421). The external AUC therefore depended on the analysis population: it was high in the single-label four-category subset but lower for several categories when all images, including multi-label images, were scored, and the difference between the internal and external four-category AUCs (0.918 versus 0.939) should not be read as better discrimination on RFMiD. Of the single-label RFMiD images, 128 were assigned to cataract, a category not annotated in RFMiD; 121 of them carried media haze (MH) as their only RFMiD label.

Their cataract status could not be verified, and these predictions may partly reflect opacity of the ocular media rather than unrelated errors. At 224 pixels, models trained on propagated labels had a lower macro AUC on RFMiD than models trained on eye-level labels (0.892 versus 0.926; difference −0.034, 95% CI −0.050 to −0.019) but a higher macro-F1 (0.667 versus 0.632; difference 0.035, 0.004 to 0.070) and balanced accuracy (0.608 versus 0.572), so the two label sources differed in opposite directions for ranking and for threshold-dependent metrics.

Table 6. External validation of the frozen models on RFMiD (network weights not updated). (**a**) Four diagnostic categories shared by both datasets (normal, diabetic retinopathy, AMD, myopia): ensemble performance [95% CI] on the ODIR test set restricted to these classes (681 images; patient-level bootstrap) and on single-label RFMiD images of these classes (1214 images; image-level bootstrap). Secondary class sets are in Supplementary Table S10, and the bootstrap outputs for all class sets in Supplementary Data S2. (**b**) One-versus-rest AUC [95% CI, image-level bootstrap] on all 3200 RFMiD images, including multi-label images; *n* = images positive for the label.

(**a**) Four shared classes

Model	Input (px)	Macro-F1, ODIR	Macro-F1, RFMiD	Balanced Accuracy, ODIR	Balanced Accuracy, RFMiD	Macro AUC, ODIR	Macro AUC, RFMiD
EfficientNetB0	384	0.815 [0.769–0.852]	0.663 [0.617–0.707]	0.764 [0.705–0.820]	0.612 [0.571–0.654]	0.918 [0.896–0.937]	0.939 [0.925–0.953]
EfficientNetB0	224	0.710 [0.650–0.758]	0.632 [0.585–0.675]	0.676 [0.610–0.741]	0.572 [0.533–0.611]	0.887 [0.863–0.907]	0.926 [0.912–0.939]
DenseNet121	384	0.769 [0.711–0.812]	0.610 [0.561–0.655]	0.729 [0.664–0.787]	0.568 [0.527–0.611]	0.919 [0.896–0.938]	0.929 [0.911–0.946]
DenseNet121	224	0.697 [0.640–0.742]	0.639 [0.594–0.680]	0.665 [0.600–0.723]	0.593 [0.553–0.637]	0.867 [0.836–0.892]	0.900 [0.883–0.916]

(**b**) One-versus-rest AUC, all 3200 RFMiD images

Model	Input (px)	Normal (*n* = 669)	Diabetic Retinopathy (*n* = 632)	AMD (*n* = 169)	Myopia (*n* = 167)	Optic Disc Cupping (*n* = 445)
EfficientNetB0	384	0.796 [0.779–0.811]	0.917 [0.903–0.930]	0.863 [0.828–0.894]	0.857 [0.823–0.888]	0.748 [0.720–0.774]
EfficientNetB0	224	0.785 [0.768–0.801]	0.881 [0.864–0.897]	0.894 [0.870–0.916]	0.877 [0.847–0.903]	0.749 [0.722–0.774]
DenseNet121	384	0.760 [0.742–0.778]	0.896 [0.881–0.911]	0.861 [0.828–0.891]	0.870 [0.840–0.898]	0.693 [0.661–0.723]
DenseNet121	224	0.795 [0.781–0.811]	0.830 [0.810–0.850]	0.886 [0.856–0.913]	0.805 [0.766–0.844]	0.731 [0.703–0.758]

The resolution effect transferred for diabetic retinopathy (AUC 0.881 to 0.917 for EfficientNetB0; 0.830 to 0.896 for DenseNet121) but not for AMD or myopia, and for EfficientNetB0, the mean AUC over the mapped classes did not differ between 384 and 224 pixels (difference −0.001, $p = 0.81$). At 384 pixels, EfficientNetB0 exceeded DenseNet121 externally (difference 0.020, 95% CI 0.010–0.030, $p < 0.001$).

3.9. Excluded Multi-Disease Images

For the 356 excluded ODIR images, the top-ranked prediction matched one of the diagnoses in 69.1% (chance 25.1%; Supplementary Table S7 and Data S4). The top-two predictions contained at least one diagnosis in 89.0%, but all diagnoses in only 21.1%, and 23.6% were predicted normal. These images were never used for training, model selection or calibration, but for 166 of them, the fellow eye was in the development set (37 in the hold-out test set; 153 had no single-label fellow eye); this is therefore an exploratory sensitivity analysis rather than an evaluation on unseen patients. Top-1 agreement was 72.3% (95% CI 65.7–78.9%) when the fellow eye was in the development set, 69.3% (60.9–77.3%) when there was no single-label fellow eye and 54.1% (37.8–70.3%) when the fellow eye was in the hold-out test set (Supplementary Table S14); the difference between the first two groups, of 3.0 percentage points (−7.5 to 13.6), did not establish a difference. The wide interval does not exclude inflation, and the groups differed in disease combinations. A development-set eye was in the training partition of four of the five fold models and in the validation partition of the fifth.

3.10. Local Recalibration

Table 7 uses the official RFMiD test (249 images) and validation (244 images) splits, whereas Table 6 pools all 1214 single-label four-category images across splits; the unrecalibrated test-split macro-F1 (0.657) therefore differs slightly from the pooled value (0.663). Without recalibration, the EfficientNetB0 (384 pixels) ensemble reached a macro-F1 of 0.657 (95% CI 0.552–0.741) on the 249 shared-class images of the RFMiD test split (Table 7). Restricting predictions to the four evaluated categories changed this little (0.670), whereas offsets fitted on the 721 training-split images raised macro-F1 to 0.735 (0.635–0.810) and balanced accuracy from 0.607 to 0.681, with a small change in macro AUC (0.927 to 0.934), whose paired confidence interval included zero. The validation split showed the same pattern (0.644 to 0.714), as did DenseNet121 (0.624 to 0.752) and EfficientNetV2-S (0.661 to 0.765; Supplementary Table S11; per-image predictions before and after recalibration are given in Supplementary Data S3). With 100 local images, macro-F1 reached 0.701 ± 0.029, about half of the full gain, and with 400 images, 0.735 ± 0.011 (Figure 3; Supplementary Table S12); 25 images gave no meaningful improvement (0.669 ± 0.029). The individual intervals overlap because the test split is small; overlapping intervals do not indicate the absence of an effect, and the direction was consistent across models, splits and subsample sizes. On the same bootstrap resamples, the paired differences between offset fitting and no recalibration on the test split were +0.079 for macro-F1 (95% CI 0.006 to 0.162; $p = 0.034$), +0.074 for balanced accuracy (0.009 to 0.149; $p = 0.023$) and +0.008 for macro AUC (−0.003 to 0.019; $p = 0.16$); on the validation split, the macro-F1 difference was +0.070 (0.001 to 0.141), in the same direction. By category, on the test split (Supplementary Table S13), offsets raised sensitivity for normal images from 0.709 to 0.925, for AMD from 0.500 to 0.611 and for myopia from 0.294 to 0.412, but lowered it for diabetic retinopathy from 0.925 to 0.775, while diabetic retinopathy precision rose from 0.622 to 0.849; the validation split showed the same trade-off (diabetic retinopathy sensitivity 0.975 to 0.914). When the offsets were constrained to keep diabetic retinopathy sensitivity at or above 0.90 on the training split, test macro-F1 was 0.748 with a diabetic retinopathy sensitivity of 0.825 (95% CI 0.740–0.906); a constraint of 0.95 gave 0.699 and 0.863 (0.783–0.934). Both targets were met on the validation split (macro-F1 0.739 with sensitivity 0.914; 0.703 with 0.963; Supplementary Table S19) but not on the 80 diabetic retinopathy images of the test split, where the interval included the 0.90 target but not the 0.95 target. The 0.90-constrained offsets also reached a slightly higher training-split macro-F1 than the unconstrained fit (0.791 versus 0.789), so the unconstrained coordinate ascent stopped at a local optimum and, at this target, the constraint did not cost macro-F1; the higher test macro-F1 under the constraint (0.748 versus 0.735) is therefore not evidence that the constraint itself improves macro-F1. Differences were computed from unrounded values and may therefore differ by 0.001 from differences between the rounded entries in Table 7. Macro-F1 was the primary metric; the balanced-accuracy and AUC tests are exploratory, and none of these tests was corrected for multiplicity.

Table 7. Local recalibration on RFMiD (network weights not updated). Offsets fitted on the official training split (721 shared-class images); evaluation on the official test split (249) and validation split (244). Values with 95% CI (image-level bootstrap). Restriction = predictions limited to the four evaluated categories; offsets = class-wise log-probability offsets maximising macro-F1; macro AUC computed from the offset-adjusted, renormalised probabilities. Table 6 pools all 1214 single-label four-category images across splits, so its unrecalibrated values differ slightly from the test-split values here.

Model	Method	Macro-F1, Test	Balanced Accuracy, Test	Macro AUC, Test	Macro-F1, Validation
EfficientNetB0 384	None	0.657 [0.552–0.741]	0.607 [0.522–0.690]	0.927 [0.890–0.956]	0.644 [0.529–0.736]
EfficientNetB0 384	Restriction	0.670 [0.571–0.755]	0.628 [0.543–0.712]	0.930 [0.894–0.958]	0.683 [0.567–0.767]

EfficientNetB0 384	Offsets	0.735 [0.635–0.810]	0.681 [0.588–0.767]	0.934 [0.901–0.963]	0.714 [0.611–0.796]
EfficientNetB0 384	Restriction + offsets	0.721 [0.619–0.801]	0.678 [0.586–0.766]	0.934 [0.902–0.963]	0.723 [0.622–0.802]
DenseNet121 384	None → offsets	0.624 → 0.752	0.580 → 0.730	0.909 → 0.927	0.549 → 0.719
EfficientNetV2-S 384	None → offsets	0.661 → 0.765	0.637 → 0.721	0.947 → 0.956	0.528 → 0.753

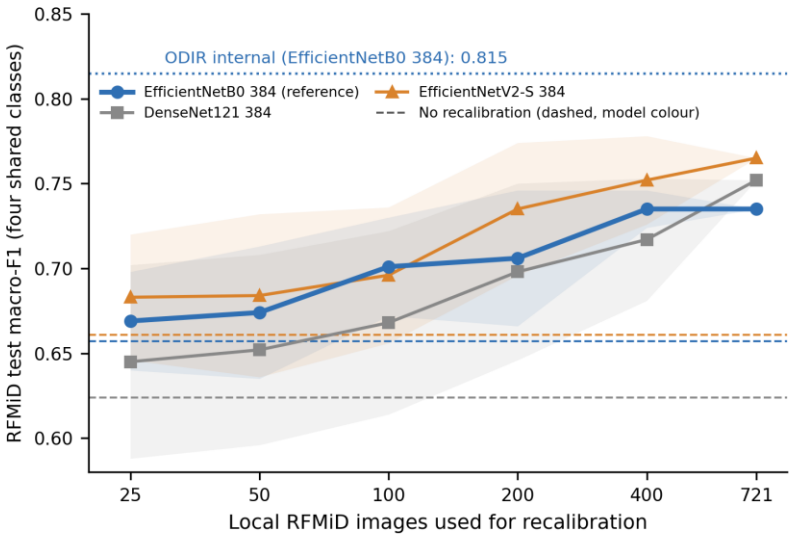


Figure 3. Local data needed for recalibration on RFMiD. Test-split macro-F1 (249 shared-class images) after fitting class-wise offsets on stratified random subsets of the official training split (mean ± SD over 20 repeats; 721 = full split, single fit). Dashed lines: No recalibration; dotted line: ODIR internal test performance of EfficientNetB0 (384 pixels) on the same four classes (0.815). The SD reflects variation across calibration subsets conditional on the fixed RFMiD test set, rather than uncertainty across new sites. Shaded bands: ±1 SD, in the colour of each model.

3.11. Performance by Sex and Age

On the hold-out test set, the EfficientNetB0 (384 pixels) ensemble reached a macro-F1 of 0.635 (95% CI 0.553–0.697) in women (385 images) and 0.669 (0.601–0.725) in men (471 images), and 0.671 (0.589–0.730), 0.630 (0.548–0.691) and 0.644 (0.566–0.700) in patients aged <50, 50–64 and ≥65 years; all intervals overlapped, and macro AUC ranged from 0.896 to 0.940 (Supplementary Table S17). The <50-year group contained no cataract images, so its metrics cover seven classes.

4. Discussion

In our preliminary pipeline, near-perfect accuracy did not reflect near-perfect disease recognition. The preliminary model had been trained and evaluated on the same images, which accounts for almost all of the gap between its 98.2% and its held-out performance. In the controlled experiment with fixed evaluation eyes and matched training-set size, adding the fellow eye to training raised macro-F1 in all three seeds (by 0.032 for the three-seed ensembles, 95% CI −0.004 to 0.068), mainly through accuracy (difference 0.031, 0.003 to 0.059) rather than discrimination. Because the fellow-eye training sets also contained about 340 more patients, part of this gain may reflect patient diversity rather than leakage, whereas the larger accuracy gain for eyes whose fellow eye shared the diagnosis (0.039 versus 0.009) is what reuse of the fellow eye's label would produce. The inflation from fellow-eye leakage is therefore modest and, for macro-F1, statistically uncertain; in accuracy (0.031), it is an order of magnitude smaller than the gap produced by training-set evaluation (1.000 versus 0.532) and much weaker than the leakage between

neighbouring slices reported in volumetric imaging [14]; patient-level splitting nevertheless remains necessary, because this inflation is of the size of the differences between architectures reported in Section 3.3, and its upper confidence limit (0.068) approaches the resolution gain in Section 3.4.

Label propagation affected discrimination. On identical images, folds and test sets, models trained on propagated labels had a lower macro AUC against eye-level diagnoses (0.877 versus 0.897) and also on RFMiD's specialist labels (0.892 versus 0.926), whereas threshold-dependent metrics did not consistently favour either label source: the macro-F1 difference on ODIR-5K was not significant, and on RFMiD, the propagated-label models had a higher macro-F1, possibly because training on propagated labels shifted their operating point in a direction that happened to suit RFMiD's class mix. This is the same separation of ranking from operating point that we observed externally. A metric computed against propagated labels measures agreement with a reference that, for 16.4% of images, differs from the eye's own annotation. The eye-level labels are derived from the provider's consensus grading by ophthalmologists, and our mapping reproduced the provider's own category assignments; for ODIR-5K image-level tasks, the per-eye keywords therefore describe the photographed eye more directly than the patient flags. Studies that use the patient flags should state how labels were mapped to images and should split at the patient level, in line with current reporting guidance for medical imaging AI [32]; the completed checklist for this study is given in Supplementary Table S22.

Under controlled conditions, the CNNs were closer than headline figures suggest: EfficientNetB0 led in macro AUC at 224 pixels, but threshold-dependent metrics did not separate the models, and at 384 pixels, no statistically significant difference was detected among EfficientNetB0, DenseNet121 and EfficientNetV2-S. The practical argument for EfficientNetB0 is its smaller size and, in our environment, shorter epochs. In these experiments, input resolution mattered more than architecture. Whether this gain continues beyond 384 pixels was not tested. The gain at 384 pixels appeared in both architectures and was concentrated in diabetic retinopathy and AMD, whose diagnostic signs (microaneurysms, haemorrhages, exudates, drusen) occupy only a few pixels at 224. Cataract and myopia, recognised from global appearance, did not benefit.

External validation separated discrimination from operating point. In the single-label subset of the four shared diagnostic categories, macro AUC remained high on RFMiD, whereas macro-F1 and balanced accuracy fell by about 0.15; across all 3200 images, including multi-label images, one-versus-rest AUCs were lower (0.857–0.917 for the mapped disease categories and 0.796 for normal), so the level of discrimination depends on the analysis population. RFMiD differs in cameras, population, class mix and labelling scheme, and a softmax trained on ODIR-5K places its decision boundaries for the ODIR distribution. The 128 images assigned to cataract, 121 of which were labelled media haze in RFMiD, illustrate the mismatch between the two label systems rather than confirmed errors, because RFMiD does not annotate cataract. These results suggest that decision thresholds may require local recalibration before deployment even when AUC remains high. AUC was likewise proved to be far less sensitive than macro-F1 when evaluation conditions change, as in a pooled-data comparison of image-level and pseudo-patient-level splits [11]. Clinically, a model validated by AUC alone could therefore appear ready for deployment while its per-image decisions are unreliable at a new site; on RFMiD, balanced accuracy fell from 0.764 to 0.612. Only the diabetic retinopathy gain from higher resolution replicated externally, which supports the lesion-scale explanation for that class but argues against generalising the 384-pixel advantage to all diseases. The excluded multi-disease images were not unrecognisable: the model usually identified one diagnosis but rarely both, a pattern to which the single-label formulation may have contributed.

The operating-point shift was partly correctable in this dataset. On the same RFMiD test split, offsets fitted on the 721 labelled training-split images improved macro-F1 by 0.079 (95% CI 0.006–0.162) without updating network weights, and 100 images provided about half of this gain, whereas restricting predictions to the four evaluated categories changed macro-F1 little (0.657 to 0.670). This supports the potential value of a small labelled calibration set from the target site, although deployment and clinical utility were not evaluated. EfficientNetV2-S, larger and more recent, changed internal performance little and did not improve external macro-F1 (0.638 versus 0.663), suggesting that architecture was not the limiting factor.

Recalibration also shows the clinical cost of optimising a summary metric. Because macro-F1 weights categories equally, offset fitting reduced diabetic retinopathy false-positive predictions from 45 to 11 at the cost of sensitivity (74/80 to 62/80; false negatives 6 to 18), while correct predictions for normal images increased from 95/134 to 124/134. In screening, a missed case of referable disease costs more than a false referral, so this trade-off would be unacceptable without explicit control. Constraining the offsets to a minimum diabetic retinopathy sensitivity recovered part of it (test sensitivity 0.825 (95% CI 0.740–0.906) for a training-split target of 0.90 and 0.863 (0.783–0.934) for 0.95); both targets were met on the validation split but not on the test split, which contained only 80 diabetic retinopathy images. At the 0.90 target, the constrained offsets had a macro-F1 no lower than that of the pre-specified fit (M2) on the training, validation and test splits (0.791 versus 0.789, 0.739 versus 0.714 and 0.748 versus 0.735) and a diabetic retinopathy sensitivity that was higher on the training and test splits (0.925 versus 0.896 and 0.825 versus 0.775) and equal on the validation split (0.914 in both), so part of the sensitivity lost with M2 reflects the pre-specified fitting procedure stopping at a local optimum rather than the macro-F1 objective alone; we nevertheless report M2 as the primary result because its procedure was fixed before computation, and even with the constraint, test sensitivity remained below its value without recalibration (0.825 versus 0.925). A target met on calibration data may therefore not hold at evaluation, the number of referable cases limits how precisely it can be verified, and the recalibration objective should encode the intended clinical operating point rather than macro-F1 alone.

Several commonly proposed refinements had no measurable benefit. Class weighting traded overall accuracy for balanced accuracy without changing macro-F1, and decision-offset adjustment on out-of-fold predictions did not provide a consistent benefit. For hypertensive retinopathy, despite a test AUC of 0.861, sensitivity under the argmax decision rule was only 0.133 (2 of 15 images). Its signs, arteriolar narrowing and arteriovenous nicking [33], are subtle vascular changes, and the limited number of development images (89, fewer in each training fold) may have contributed to this result; more data or explicit vessel-level features, for example, from open retinal vessel-analysis tools, may be needed.

The numerical failures with TF32 disabled began in the forward pass of frozen pretrained layers during the first epoch (Supplementary Text S1). They may be related to the unsupported GPU and software combination. In a separate study that used the same patient-level splits (manuscript in preparation), the 224-pixel benchmark of the four CNNs was repeated on a natively supported software stack (NVIDIA NGC TensorFlow 25.02 container; TensorFlow 2.17, CUDA 12.8); none of 88 training runs produced non-finite losses, whereas in the configuration used here, 15 of 16 FP32 runs (TF32 disabled) ended with non-finite losses (DenseNet121 7/7, ResNet50V2 7/7 and EfficientNetB0 1/2; Supplementary Table S24). This supports, but does not prove, an environmental cause; a full analysis of software stack and arithmetic precision will be reported separately. Timing results, including the absence of a TF32 speed advantage, are specific to this environment and should not be generalised to model rankings.

Grad-CAM maps were illustrative only: one glaucoma map highlighted the image border, the 12 × 12 feature map localises coarsely, and without expert annotation, such maps cannot support claims of clinical concordance [34].

Limitations

External validation covered only the four of ODIR-5K's eight categories that are shared by both datasets: cataract and hypertensive retinopathy could not be evaluated, glaucoma was approximated by optic disc cupping, and RFMiD lacks patient identifiers, so its bootstrap intervals assume independent images. Excluding 356 multi-disease images simplified the task but departs from clinical case mix, so the results should not be read as multi-disease diagnostic performance, and a multi-label formulation is a natural extension (Section 3.9). The labels rest on the provider's consensus grading [5]; our mapping reproduced the provider's categories, but no images were re-graded for this study, the provider did not report inter-grader agreement, and the duplicate review found one pair of identical eyes with different categories, so the 16.4% figure measures disagreement with the provider's eye-level annotation rather than with an independent reading. Duplicate screening was restricted post hoc to bilateral patient matches and may have missed duplicates involving a single eye. The hold-out test set contained few images in minority classes (15 hypertension, 30 glaucoma), and their class-specific intervals are wide. The leakage reproduction was run once; the descriptive label-by-split conditions differ in the image sets and test sets, and the controlled experiments that address this used one architecture and one resolution, with three seeds for the fellow-eye experiment and one for the label-source experiment. In the fellow-eye experiment, training-set size and class counts were matched but the number of distinct patients was not (333–341 more patients per fold with fellow eyes), and we did not rerun it with matched patient numbers, so the 0.032 gain may partly reflect patient diversity rather than leakage. Recalibration was evaluated on splits of the same external dataset, so a genuinely new site may need more local data, and its 249-image test split gives wide intervals. Subgroup analyses by sex and age were descriptive and had limited precision, and RFMiD provides no demographic data. Only five CNN architectures and two resolutions were examined: training EfficientNetB0 at the native 512-pixel resolution with the protocol's batch size exceeded the memory of the 16 GB GPU in three attempts, and we did not train with a smaller batch because the comparison with 384 pixels would then confound resolution with batch size; gradient accumulation, which would avoid this confound, requires a custom training loop that we did not implement. ConvNeXt-Tiny could not be trained here, EfficientNetV2-S used a smaller batch size, so its comparison with EfficientNetB0 is confounded by batch size, and frozen retinal foundation-model features [35] were not evaluated. All experiments ran on one unsupported GPU and software configuration, which limits the generality of the timing and numerical findings; evidence from a supported configuration is so far limited to the separate study summarised in Section 4.

5. Conclusions

In our preliminary ODIR-5K pipeline, near-perfect accuracy came from evaluating the model on the images it had been trained on; with test eyes held fixed, training on the patients' fellow eyes raised macro-F1 by 0.032 (95% CI −0.004 to 0.068; possibly less, because the training sets also differed in patient numbers), far less than training-set evaluation but comparable to differences between architectures, and models trained on patient-level diagnoses propagated to individual eyes discriminated less well on both internal and external data. With eye-level labels, patient-level splits and validation-only model selection, EfficientNetB0 at 384 pixels reached a macro-F1 of 0.659, an accuracy of 70.8% and a macro AUC of 0.913 on held-out patients. On an external dataset, accuracy

for the four shared diagnostic categories was maintained (0.749 internally and 0.753 externally) and discrimination remained high in the evaluated single-label subset (macro AUC 0.939), but macro-F1 fell (0.815 to 0.663), and offsets fitted on labelled local images improved macro-F1 on the RFMiD test split (0.657 to 0.735) while reducing diabetic retinopathy sensitivity (0.925 to 0.775), demonstrating an operating-point trade-off; cataract and hypertensive retinopathy could not be evaluated externally, and only the diabetic retinopathy benefit of higher resolution replicated. Minority classes such as hypertensive retinopathy remain unsolved. We release the split files, label dictionary, external mapping protocol and code so that future work can be compared on the same footing.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/diagnostics16233957/s1, Text S1: Numerical diagnosis; Text S2: RFMiD mapping protocol (fixed before inference); Text S3: Experimental grid; Table S1: Keyword-to-class dictionary and exclusion rules; Table S2: Eye-level labels versus the distributor's per-image label (first keyword only); Table S3: Runs with TF32 disabled, MobileNetV2 learning-rate runs and their TF32-enabled comparators; Table S4: Threshold calibration: Nested out-of-fold estimate and test results; Table S5: Paired bootstrap comparisons between models at 224 pixels (test ensembles); Table S6: RFMiD one-versus-rest AUC with image-level bootstrap 95% CI; Table S7: Excluded multi-disease ODIR images (sensitivity analysis); Table S8: RFMiD analysis populations; Table S9: Complete RFMiD label mapping; Table S10: ODIR test versus RFMiD for all predefined class sets (ensembles, 95% CI); Table S11: Local recalibration on RFMiD, all models (test n = 249, validation n = 244); Table S12: Local data needed: RFMiD test macro-F1 after offset fitting on subsets of the training split; Table S13: Per-category effect of local recalibration on RFMiD; Table S14: Excluded multi-disease ODIR images by location of the fellow eye; Table S15: Demographics of the development and test sets; Table S16: Class-specific test performance with 95% CI (EfficientNetB0, 384 pixels); Table S17: Test performance by sex and age group; Table S18: Keyword mapping versus the provider's patient-level categories, decomposition of the 16.4% disagreement and alternative mappings; Table S19: Recalibration with a minimum diabetic retinopathy sensitivity; Table S20: Seed-wise five-fold ensembles and individual models (EfficientNetB0, 384 pixels); Table S21: Controlled experiments on label source and fellow-eye leakage; Table S22: CLAIM 2024 checklist; Table S23: Duplicate-image review; Table S24: Non-finite training losses on two software stacks in a separate study (run counts only); Data S1: run_metrics.csv; Data S2: Supplementary_Data_rfmid_ci.md; Data S3: Per-image RFMiD four-category labels, ensemble probabilities and predictions before and after recalibration; Data S4: Per-image diagnoses, predictions, patient identifiers and fellow-eye group for the excluded multi-disease ODIR images; Data S5: Per-run metrics of the revision experiments (format of Data S1, with experiment, arm and evaluation set); Data S6: Numbers added in the second revision (script and outputs computed from the released predictions) and the training script and log of the preliminary analysis.

Author Contributions: Conceptualisation, H.R.C. and W.S.K.; methodology, H.R.C. and W.S.K.; software, H.R.C.; validation, W.S.K. and C.-S.L.; formal analysis, H.R.C.; investigation, H.R.C. and C.-S.L.; data curation, H.R.C.; writing—original draft preparation, H.R.C.; writing—review and editing, W.S.K. and C.-S.L.; visualisation, H.R.C.; supervision, W.S.K.; project administration, W.S.K.; funding acquisition, W.S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Regional Innovation System for Talent Development (ANCHOR) programme, funded by the Ministry of Education and Daejeon Metropolitan City through the Daejeon Anchor Centre (2026-ANCHOR-06-001).

Institutional Review Board Statement: This study is a secondary analysis of two publicly released, de-identified image datasets (ODIR-5K and RFMiD) that their providers made available for research. The authors had no access to personally identifiable information, had no contact or interaction with the individuals concerned, and collected no new data. Research of this kind—using

publicly available information without collecting or recording personally identifiable information—falls within the categories eligible for exemption from review under the Bioethics and Safety Act of the Republic of Korea (Article 13 of the Enforcement Rule of the Act). Ethical review and approval were therefore not sought for this study.

Informed Consent Statement: Informed consent was not obtained by the authors because this study used only publicly available, de-identified data and the authors had no contact with the individuals concerned.

Data Availability Statement: The ODIR-5K images and annotations [5] were obtained on 16 April 2026 from the Kaggle copy of the ODIR-2019 Grand Challenge data ("Ocular Disease Recognition", https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k; accessed on 16 April 2026), using its full_df.csv annotation file and preprocessed_images (512 × 512 pixels). The original source (https://odir2019.grand-challenge.org/, accessed on 15 October 2026) describes the data as de-identified and published under the ethical and privacy rules of China, and its dataset page (accessed on 15 October 2026) does not state an explicit licence; the licence shown on the Kaggle copy is not used as a basis here. The RFMiD images and labels [19] were obtained on 7 October 2026 from a Kaggle copy ("Retinal Disease Classification", https://www.kaggle.com/datasets/andrewmvd/retinal-disease-classification; accessed on 7 October 2026) of the original release (IEEE DataPort, https://doi.org/10.21227/s3g7-st65) and were used under the CC BY 4.0 licence stated by the dataset authors. Neither image set is redistributed here: the repository contains only code, patient-level split lists, the keyword-to-class dictionary, the RFMiD mapping protocol and result files, including per-run predictions, and users must obtain the images from the original sources under their terms; the images derived from them, such as the Grad-CAM maps, and model weights are not redistributed either. These materials, together with the revision analyses and their protocol and deviation records, are archived at https://doi.org/10.5281/zenodo.24213394 (version 1.1.1-revision1).

Acknowledgments: During the preparation of this study, the authors used Claude (Anthropic; Claude Code, Claude Fable 5.2, 2026) for the purposes described in Section 2.9. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AMD	Age-related macular degeneration
AUC	Area under the receiver operating characteristic curve
CI	Confidence interval
CNN	Convolutional neural network
Grad-CAM	Gradient-weighted class activation mapping
ODIR-5K	Ocular Disease Intelligent Recognition dataset
RFMiD	Retinal Fundus Multi-disease Image Dataset
TF32	TensorFloat-32

References

1. Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M.C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **2016**, *316*, 2402–2410. https://doi.org/10.1001/jama.2016.17216.
2. Ting, D.S.W.; Cheung, C.Y.L.; Lim, G.; Tan, G.S.W.; Quang, N.D.; Gan, A.; Hamzah, H.; Garcia-Franco, R.; Yeo, I.Y.S.; Lee, S.Y.; et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* **2017**, *318*, 2211–2223. https://doi.org/10.1001/jama.2017.18152.
3. Ting, D.S.W.; Pasquale, L.R.; Peng, L.; Campbell, J.P.; Lee, A.Y.; Raman, R.; Tan, G.S.W.; Schmetterer, L.; Keane, P.A.; Wong, T.Y. Artificial intelligence and deep learning in ophthalmology. *Br. J. Ophthalmol.* **2019**, *103*, 167–175. https://doi.org/10.1136/bjophthalmol-2018-313173.
4. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.W.M.; van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. https://doi.org/10.1016/j.media.2017.07.005.
5. Li, N.; Li, T.; Hu, C.; Wang, K.; Kang, H. A benchmark of ocular disease intelligent recognition: One shot for multi-disease detection. In *Benchmarking, Measuring, and Optimizing (Bench 2020)*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2021; Volume 12614, pp. 177–193. https://doi.org/10.1007/978-3-030-71058-3_11.
6. Gour, N.; Khanna, P. Multi-class multi-label ophthalmological disease detection using transfer learning based convolutional neural network. *Biomed. Signal Process. Control* **2021**, *66*, 102329. https://doi.org/10.1016/j.bspc.2020.102329.
7. Sivaz, O.; Aykut, M. Combining EfficientNet with ML-Decoder classification head for multi-label retinal disease classification. *Neural Comput. Appl.* **2024**, *36*, 14251–14261. https://doi.org/10.1007/s00521-024-09820-w.
8. Kılıç, Ş. HybridVisionNet: An advanced hybrid deep learning framework for automated multi-class ocular disease diagnosis using fundus imaging. *Ain Shams Eng. J.* **2025**, *16*, 103594. https://doi.org/10.1016/j.asej.2025.103594.
9. Bhati, A.; Gour, N.; Khanna, P.; Ojha, A. Discriminative kernel convolution network for multi-label ophthalmic disease detection on imbalanced fundus image dataset. *Comput. Biol. Med.* **2023**, *153*, 106519. https://doi.org/10.1016/j.compbiomed.2022.106519.
10. Sah, U.K.; Chatterjee, J.M.; Sujatha, R. Multi-class eye disease classification using deep learning EfficientNetB0 fusion techniques. *Sci. Rep.* **2026**, *16*, 6368. https://doi.org/10.1038/s41598-026-35357-0.
11. Vasilev, R.; Savchenko, A.; Blinov, P.; Svetina, T.; Kudin, S.; Romanenko, N.; Sarana, Y.; Khizhnyak, G.; Demchinsky, A.; Shcheglova, T. Artificial intelligence framework for multi-pathology risk assessment from retinal fundus images: Deep learning approach to 15-disease screening. *Front. Med.* **2026**, *13*, 1778404. https://doi.org/10.3389/fmed.2026.1778404.
12. Kaufman, S.; Rosset, S.; Perlich, C.; Stitelman, O. Leakage in data mining: Formulation, detection, and avoidance. *ACM Trans. Knowl. Discov. Data* **2012**, *6*, 15. https://doi.org/10.1145/2382577.2382579.
13. Kapoor, S.; Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **2023**, *4*, 100804. https://doi.org/10.1016/j.patter.2023.100804.
14. Yagis, E.; Atnafu, S.W.; García Seco de Herrera, A.; Marzi, C.; Scheda, R.; Giannelli, M.; Tessa, C.; Citi, L.; Diciotti, S. Effect of data leakage in brain MRI classification using 2D convolutional neural networks. *Sci. Rep.* **2021**, *11*, 22544. https://doi.org/10.1038/s41598-021-01681-w.
15. Zech, J.R.; Badgeley, M.A.; Liu, M.; Costa, A.B.; Titano, J.J.; Oermann, E.K. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med.* **2018**, *15*, e1002683. https://doi.org/10.1371/journal.pmed.1002683.
16. DeGrave, A.J.; Janizek, J.D.; Lee, S.I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat. Mach. Intell.* **2021**, *3*, 610–619. https://doi.org/10.1038/s42256-021-00338-7.
17. Roberts, M.; Driggs, D.; Thorpe, M.; Gilbey, J.; Yeung, M.; Ursprung, S.; Aviles-Rivero, A.I.; Etmann, C.; McCague, C.; Beer, L.; et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* **2021**, *3*, 199–217. https://doi.org/10.1038/s42256-021-00307-0.
18. Varoquaux, G.; Cheplygina, V. Machine learning for medical imaging: Methodological failures and recommendations for the future. *NPJ Digit. Med.* **2022**, *5*, 48. https://doi.org/10.1038/s41746-022-00592-y.
19. Pachade, S.; Porwal, P.; Thulkar, D.; Kokare, M.; Deshmukh, G.; Sahasrabuddhe, V.; Giancardo, L.; Quellec, G.; Mériaudeau, F. Retinal Fundus Multi-Disease Image Dataset (RFMiD): A dataset for multi-disease detection research. *Data* **2021**, *6*, 14. https://doi.org/10.3390/data6020014.

20. Tan, M.; Le, Q.V. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*; PMLR: Cambridge, MA, USA, 2019; Volume 97, pp. 6105–6114. Available online: https://proceedings.mlr.press/v97/tan19a.html (accessed on 20 November 2026).
21. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2017; pp. 2261–2269. https://doi.org/10.1109/CVPR.2017.243.
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In *Computer Vision—ECCV 2016*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2016; Volume 9908, pp. 630–645. https://doi.org/10.1007/978-3-319-46493-0_38.
23. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2018; pp. 4510–4520. https://doi.org/10.1109/CVPR.2018.00474.
24. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2009; pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848.
25. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
26. Buda, M.; Maki, A.; Mazurowski, M.A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.* **2018**, *106*, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011.
27. Johnson, J.M.; Khoshgoftaar, T.M. Survey on deep learning with class imbalance. *J. Big Data* **2019**, *6*, 27. https://doi.org/10.1186/s40537-019-0192-5.
28. Choquette, J.; Gandhi, W.; Giroux, O.; Stam, N.; Krashinsky, R. NVIDIA A100 Tensor Core GPU: Performance and innovation. *IEEE Micro* **2021**, *41*, 29–35. https://doi.org/10.1109/MM.2021.3061394.
29. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2017; pp. 618–626. https://doi.org/10.1109/ICCV.2017.74.
30. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; Chapman & Hall: New York, NY, USA, 1993. https://doi.org/10.1201/9780429246593.
31. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
32. Tejani, A.S.; Klontzas, M.E.; Gatti, A.A.; Mongan, J.T.; Moy, L.; Park, S.H.; Kahn, C.E., Jr.; CLAIM 2024 Update Panel. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. *Radiol. Artif. Intell.* **2024**, *6*, e240300. https://doi.org/10.1148/ryai.240300.
33. Wong, T.Y.; Mitchell, P. Hypertensive retinopathy. *N. Engl. J. Med.* **2004**, *351*, 2310–2317. https://doi.org/10.1056/NEJMra032865.
34. Ghassemi, M.; Oakden-Rayner, L.; Beam, A.L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Health* **2021**, *3*, e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9.
35. Zhou, Y.; Chia, M.A.; Wagner, S.K.; Ayhan, M.S.; Williamson, D.J.; Struyven, R.R.; Liu, T.; Xu, M.; Lozano, M.G.; Woodward-Court, P.; et al. A foundation model for generalizable disease detection from retinal images. *Nature* **2023**, *622*, 156–163. https://doi.org/10.1038/s41586-023-06555-x.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.