Supplementary Materials

Calibrated Confidence and Confidence-Stratified Triage for Diabetic Retinopathy Severity Grading: A Class-Imbalance-Aware Deep Learning Study on the DDR Dataset

Roseline Oluwaseun Ogundokun, Rotimi-Williams Bello, Oluwasegun Julius Aroba, Samson Udoka Alfred

All items were produced by the revised training and evaluation notebook (Code S1) from the group-aware rerun described in Section 3.2 of the main text. Unless stated otherwise, results refer to the main model (EfficientNet-B0, class-weighted cross-entropy) and to the pooled out-of-fold predictions for all 12,522 images. Variant A is the raw softmax output, B the temperature-scaled output and C the temperature-scaled MC-Dropout output (M = 20). Confidence intervals are 95% group-clustered bootstrap intervals (1,000 resamples of the 12,359 near-duplicate groups) unless stated as Wilson intervals.

1. Configuration and data

Table S1. Software, hardware and training configuration.

Item	**Value**
GPU	NVIDIA GeForce RTX 5080 Laptop GPU
Operating system	Windows-11-10.0.26200-SP0
Python	3.14.7
CUDA / cuDNN	12.8 / 91900
torch	2.11.0+cu128
torchvision	0.26.0+cu128
numpy	2.5.3
pandas	3.0.6
scikit-learn	1.9.1
scipy	1.18.1
scikit-image	0.26.0
opencv-python	5.0.0.93
matplotlib	3.11.2
Random seed	2026
Outer folds	5
Split mode	groups of near-duplicate images, stratified by grade
Near-duplicate threshold	24 of 256 bits (pHash and dHash)
Calibration fraction	0.15
Image size	384 × 384 (ViT-B/16: 224 × 224)
Batch size	16
Epochs	12

Optimiser	AdamW, learning rate 0.0003, weight decay 0.0001, cosine schedule
Dropout (projection head)	0.3
Label smoothing	0.05
MC passes (stored / used)	100 / 20
ECE bins	15
Bootstrap resamples	1000
Triage thresholds	0.5, 0.6, 0.7, 0.8, 0.9, 0.95
Referable DR	ICDR grade ≥ 2
Calibration-split target for threshold choice	referable-DR false-clear rate ≤ 1%

Table S2. Images per grade in each partition of each fold, and the temperature fitted on each calibration split.

Fold	Partition	No DR	Mild NPDR	Moderate NPDR	Severe NPDR	Proliferative DR	Total	Temperature
1	calibration	716	72	512	27	104	1431	0.707
1	outer test	1254	126	895	47	183	2505	
1	inner training	4296	432	3070	162	626	8586	
2	calibration	716	72	512	27	104	1431	0.679
2	outer test	1253	126	895	48	183	2505	
2	inner training	4297	432	3070	161	626	8586	
3	calibration	716	72	512	27	104	1431	0.683
3	outer test	1253	126	896	47	182	2504	
3	inner training	4297	432	3069	162	627	8587	
4	calibration	716	72	512	27	104	1431	0.625
4	outer test	1253	126	895	47	183	2504	
4	inner training	4297	432	3070	162	626	8587	
5	calibration	716	72	512	27	104	1431	0.696
5	outer test	1253	126	896	47	182	2504	
5	inner training	4297	432	3069	162	627	8587	

Table S3. Near-duplicate audit (part a: summary).

Quantity	Value
Images	12522

Groups after merging near-duplicates	12359
Near-duplicate pairs	175
Images in groups of two or more	317
Largest group (images)	4
Groups containing different grades	4
Pairs split across train/test by the original image-level folds	140
Images with a near-duplicate on the other side of the original split	255
Hamming threshold (of 256 bits)	24

Table S3 (continued). Near-duplicate audit (part b: images with at least one near-duplicate, by grade).

Grade	**Images with a near-duplicate**	**Share of grade**
No DR	276	4.4%
Mild NPDR	5	0.79%
Moderate NPDR	23	0.51%
Severe NPDR	0	0.0%
Proliferative DR	13	1.42%

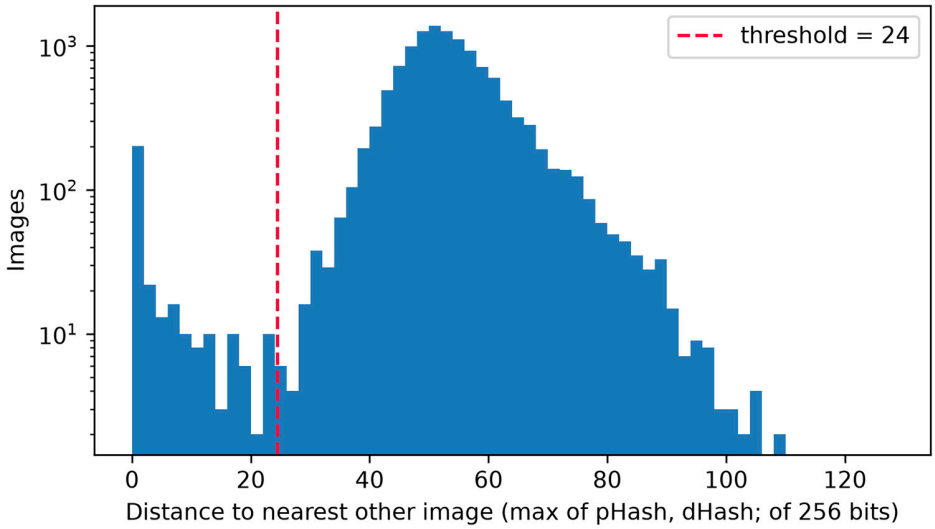


Figure S1. Distance from each image to its nearest other image (maximum of the pHash and dHash Hamming distances, of 256 bits; logarithmic count axis). A small mode of near-identical images is separated from the main distribution; the dashed line marks the threshold of 24 bits.



Figure S2. The eight matched pairs with the largest hash distance below the threshold (each row: the two images, their file names, grades and distance). These are the least similar pairs that were merged; all show the same eye.

2. Discrimination and calibration

Table S4. Fold-level metrics for the three confidence variants (CW-ECE = class-wise ECE, macro-averaged).

Variant	Fold	Accuracy	Macro-F1	Bal. acc.	MCC	ROC-AUC	PR-AUC	ECE	CW-ECE	Brier	NLL
A	1	0.8379	0.7187	0.7848	0.7511	0.9378	0.7476	0.1853	0.1128	0.3153	0.6735
A	2	0.8275	0.7002	0.7464	0.7315	0.9313	0.7479	0.1783	0.1113	0.3170	0.6656
A	3	0.8171	0.6804	0.7585	0.7179	0.9323	0.7202	0.1937	0.1157	0.3427	0.7116
A	4	0.8299	0.7113	0.7639	0.7357	0.9291	0.7607	0.1779	0.1102	0.3181	0.6686
A	5	0.8387	0.7212	0.7985	0.7535	0.9404	0.7506	0.1987	0.1159	0.3142	0.6624
B	1	0.8379	0.7187	0.7848	0.7511	0.9337	0.7456	0.0917	0.0796	0.2795	0.6340
B	2	0.8275	0.7002	0.7464	0.7315	0.9287	0.7462	0.0793	0.0760	0.2774	0.6081
B	3	0.8171	0.6804	0.7585	0.7179	0.9298	0.7180	0.0933	0.0806	0.3056	0.6755
B	4	0.8299	0.7113	0.7639	0.7357	0.9278	0.7588	0.0741	0.0686	0.2784	0.6147
B	5	0.8387	0.7212	0.7985	0.7535	0.9396	0.7491	0.0952	0.0831	0.2728	0.6070
C	1	0.8371	0.7162	0.7816	0.7497	0.9350	0.7493	0.0885	0.0806	0.2795	0.6293
C	2	0.8287	0.7020	0.7475	0.7333	0.9295	0.7471	0.0801	0.0765	0.2772	0.6046
C	3	0.8175	0.6802	0.7588	0.7188	0.9303	0.7187	0.0936	0.0815	0.3056	0.6720
C	4	0.8287	0.7080	0.7592	0.7340	0.9277	0.7572	0.0744	0.0687	0.2784	0.6103
C	5	0.8387	0.7214	0.7985	0.7536	0.9401	0.7584	0.0940	0.0840	0.2727	0.6034

Table S5. Per-grade metrics of variant C with 95% confidence intervals.

Grade	Precision	Recall	F1	Specificity	ROC-AUC	Class-wise ECE
No DR	0.934 (0.927–0.940)	0.882 (0.873–0.890)	0.907 (0.902–0.912)	0.938 (0.931–0.943)	0.973 (0.971–0.976)	0.116 (0.112–0.121)
Mild NPDR	0.296 (0.272–0.319)	0.614 (0.578–0.654)	0.399 (0.372–0.425)	0.922 (0.917–0.927)	0.888 (0.875–0.902)	0.078 (0.074–0.083)
Moderate NPDR	0.885 (0.874–0.894)	0.786 (0.773–0.797)	0.832 (0.823–0.841)	0.943 (0.937–0.948)	0.947 (0.943–0.951)	0.077 (0.073–0.083)
Severe NPDR	0.405 (0.357–0.452)	0.682 (0.621–0.741)	0.508 (0.458–0.553)	0.981 (0.978–0.983)	0.882 (0.849–0.913)	0.103 (0.101–0.105)
Proliferative DR	0.872 (0.850–0.892)	0.882 (0.861–0.902)	0.877 (0.861–0.893)	0.990 (0.988–0.992)	0.973 (0.963–0.981)	0.014 (0.013–0.017)

Table S6. Class-wise ECE (15 bins, one-vs-rest) per grade and variant, and the paired change from temperature scaling (B − A) and from MC-Dropout (C − B).

Grade	A: raw	B: temperature	C: temp. + MC	B − A	C − B
No DR	0.176 (0.172–0.181)	0.115 (0.111–0.120)	0.116 (0.112–0.121)	-0.061 (-0.062–-0.060)	0.001 (0.001–0.001)
Mild NPDR	0.096 (0.093–0.100)	0.077 (0.073–0.082)	0.078 (0.074–0.082)	-0.019 (-0.020–-0.018)	0.000 (0.000–0.001)

Moderate NPDR	0.104	(0.098–0.109)	0.076	(0.071–0.081)	0.077	(0.072–0.082)	-0.027	(-0.029–-0.026)	0.001	(0.000–0.001)
Severe NPDR	0.152	(0.150–0.154)	0.102	(0.100–0.105)	0.103	(0.101–0.106)	-0.050	(-0.050–-0.049)	0.001	(0.001–0.001)
Proliferative DR	0.035	(0.034–0.038)	0.014	(0.012–0.016)	0.014	(0.013–0.017)	-0.022	(-0.024–-0.020)	0.001	(-0.000–0.002)

Table S7. Top-label ECE under four binning schemes.

Binning	A: raw	B: temperature	C: temp. + MC
10 equal-width bins	0.1863 (0.1807–0.1922)	0.0783 (0.0748–0.0861)	0.0804 (0.0771–0.0878)
15 equal-width bins	0.1863 (0.1807–0.1923)	0.0781 (0.0746–0.0864)	0.0779 (0.0743–0.0851)
20 equal-width bins	0.1863 (0.1807–0.1922)	0.0803 (0.0764–0.0876)	0.0817 (0.0779–0.0893)
15 equal-mass bins	0.1846 (0.1789–0.1904)	0.0775 (0.0749–0.0850)	0.0804 (0.0769–0.0875)

Table S8. Top-label reliability bins of variant C (15 equal-width bins; empty bins omitted).

Bin	Images	Mean confidence	Observed accuracy	Gap (accuracy − confidence)
0.267–0.333	19	0.315	0.263	-0.052
0.333–0.400	88	0.372	0.409	+0.037
0.400–0.467	258	0.436	0.492	+0.056
0.467–0.533	428	0.501	0.516	+0.016
0.533–0.600	492	0.566	0.573	+0.007
0.600–0.667	564	0.636	0.644	+0.008
0.667–0.733	810	0.702	0.721	+0.019
0.733–0.800	1533	0.772	0.842	+0.070
0.800–0.867	4699	0.830	0.940	+0.110
0.867–0.933	2267	0.897	0.886	-0.011
0.933–1.000	1364	0.978	0.777	-0.201

3. Triage and uncertainty

Table S9. Triage operating points on the no-DR probability of variant C (n = 12,522; 5,626 referable-DR images).

τ	Auto-cleared	Referable false-clear	Wilson 95% CI	Bootstrap 95% CI	Any-DR false-clear	Share of referable missed (Wilson CI)
0.5	5,693 (45.5%)	212 (3.72%)	3.26–4.25%	3.22–4.25%	315 (5.53%)	3.77% (3.30–4.30%)
0.6	5,444 (43.5%)	170 (3.12%)	2.69–3.62%	2.65–3.62%	257 (4.72%)	3.02% (2.61–3.50%)
0.7	5,006 (40.0%)	116 (2.32%)	1.94–2.77%	1.89–2.77%	171 (3.42%)	2.06% (1.72–2.47%)
0.8	3,772 (30.1%)	55 (1.46%)	1.12–1.89%	1.07–1.88%	77 (2.04%)	0.98% (0.75–1.27%)
0.9	0 (0.0%)	0 (n/a)	n/a	n/a	0 (n/a)	0.00% (0.00–0.07%)

0.95	0 (0.0%)	0 (n/a)	n/a	n/a	0 (n/a)	0.00% (0.00–0.07%)

Table S10a. The τ = 0.80 rule applied to each outer fold.

Fold	Test images	Auto-cleared	Referable false-clear (Wilson CI)
1	2505	740 (29.5%)	10 (1.35%; 0.74–2.47%)
2	2505	831 (33.2%)	10 (1.20%; 0.65–2.20%)
3	2504	620 (24.8%)	9 (1.45%; 0.77–2.74%)
4	2504	902 (36.0%)	17 (1.88%; 1.18–3.00%)
5	2504	679 (27.1%)	9 (1.33%; 0.70–2.50%)

Table S10b. Threshold chosen on each fold's calibration split (lowest τ with a referable-DR false-clear rate ≤ 1% on that split) and applied unchanged to the fold's test images.

Fold	τ chosen on calibration split	Auto-cleared on test fold	Referable false-clear (Wilson CI)
1	0.81	592 (23.6%)	7 (1.18%; 0.57–2.42%)
2	0.81	749 (29.9%)	8 (1.07%; 0.54–2.09%)
3	0.70	941 (37.6%)	20 (2.13%; 1.38–3.26%)
4	0.85	483 (19.3%)	4 (0.83%; 0.32–2.11%)
5	0.82	282 (11.3%)	1 (0.35%; 0.06–1.98%)
Pooled		3047 (24.3%)	40 (1.31%; 0.97–1.78%)

Table S10c. Uncertainty-based auto-clear rules compared with the no-DR probability threshold at the same workload.

Rule (3,772 images cleared)	Referable false-clear	Wilson 95% CI
no-DR probability >= 0.80	55 (1.46%)	1.12–1.89%
predicted no DR, lowest MI	83 (2.20%)	1.78–2.72%
no-DR probability >= 0.70, lowest MI	74 (1.96%)	1.57–2.46%
predicted no DR, lowest entropy	52 (1.38%)	1.05–1.80%

Table S11a. Error detection and risk–coverage (500 clustered bootstrap resamples). With no abstention, the error rate is 0.170.

Uncertainty score	Error-detection AUROC	AURC	Selective error at 70% coverage
1 − maximum probability	0.650 (0.634–0.666)	0.137 (0.128–0.146)	0.102 (0.096–0.108)
Predictive entropy	0.640 (0.624–0.655)	0.138 (0.129–0.147)	0.106 (0.100–0.113)
Mutual information	0.748 (0.736–0.760)	0.093 (0.086–0.099)	0.083 (0.077–0.089)

Table S11b. Confidence and mutual information (MI) of correct and misclassified images by true grade.

True grade	Correct / errors	Median confidence (correct / errors)	Median MI (correct / errors)	MI error AUROC within grade	Mann–Whitney p
No DR	5,525 / 741	0.816 / 0.709	0.0048 / 0.0087	0.817	< 0.001
Mild NPDR	387 / 243	0.942 / 0.697	0.0057 / 0.0072	0.620	< 0.001
Moderate NPDR	3,517 / 960	0.872 / 0.776	0.0053 / 0.0081	0.675	< 0.001
Severe NPDR	161 / 75	0.984 / 0.806	0.0027 / 0.0074	0.681	< 0.001
Proliferative DR	805 / 108	0.986 / 0.778	0.0020 / 0.0109	0.870	< 0.001

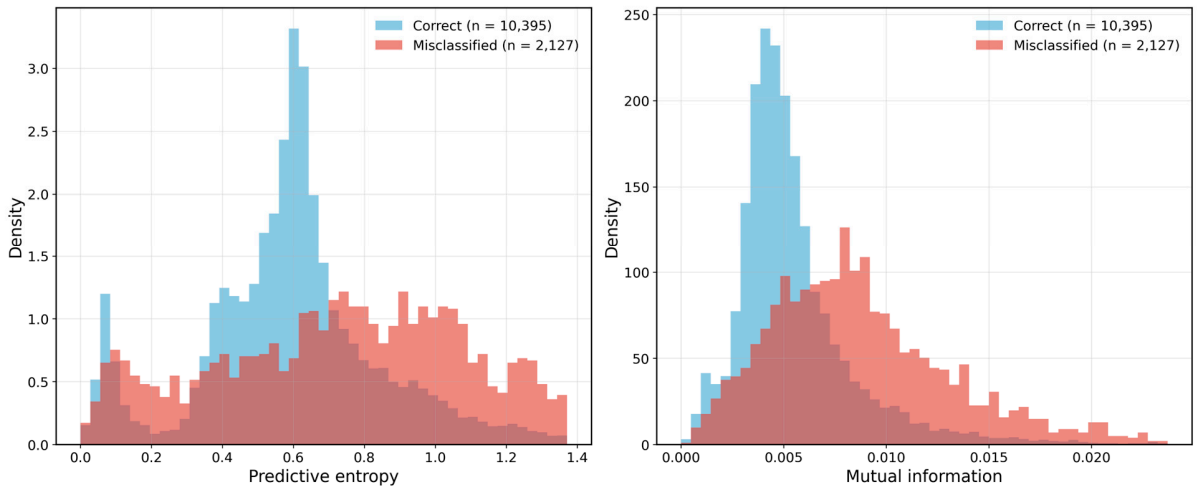


Figure S3. Distribution of MC-Dropout predictive entropy (left) and mutual information (right) for correctly classified and misclassified images.

Table S12a. Direction of errors by true grade (n / grade total, %, Wilson 95% CI).

True grade	Correct	Under-called	Referable called non-referable	Over-called	Non-referable called referable	Off by ≥ 2 grades
No DR	5525 / 6266 (88.2%; 87.4–89.0%)	0 / 6266 (0.0%; 0.0–0.1%)	0 / 6266 (0.0%; 0.0–0.1%)	741 / 6266 (11.8%; 11.0–12.6%)	240 / 6266 (3.8%; 3.4–4.3%)	240 / 6266 (3.8%; 3.4–4.3%)
Mild NPDR	387 / 630 (61.4%; 57.6–65.1%)	124 / 630 (19.7%; 16.8–23.0%)	0 / 630 (0.0%; 0.0–0.6%)	119 / 630 (18.9%; 16.0–22.1%)	119 / 630 (18.9%; 16.0–22.1%)	10 / 630 (1.6%; 0.9–2.9%)
Moderate NPDR	3517 / 4477 (78.6%; 77.3–79.7%)	673 / 4477 (15.0%; 14.0–16.1%)	673 / 4477 (15.0%; 14.0–16.1%)	287 / 4477 (6.4%; 5.7–7.2%)	0 / 4477 (0.0%; 0.0–0.1%)	349 / 4477 (7.8%; 7.0–8.6%)
Severe NPDR	161 / 236 (68.2%; 62.0–73.8%)	60 / 236 (25.4%; 20.3–31.3%)	4 / 236 (1.7%; 0.7–4.3%)	15 / 236 (6.4%; 3.9–10.2%)	0 / 236 (0.0%; 0.0–1.6%)	4 / 236 (1.7%; 0.7–4.3%)
Proliferative DR	805 / 913 (88.2%; 85.9–90.1%)	108 / 913 (11.8%; 9.9–14.1%)	11 / 913 (1.2%; 0.7–2.1%)	0 / 913 (0.0%; 0.0–0.4%)	0 / 913 (0.0%; 0.0–0.4%)	79 / 913 (8.7%; 7.0–10.7%)

Table S12b. Referable DR (grade ≥ 2) as a binary decision from the predicted grade.

Measure	Estimate (95% CI)
Sensitivity	0.878 (0.869–0.886)

Specificity	0.948 (0.943–0.953)
Positive predictive value	0.932 (0.926–0.939)
Negative predictive value	0.905 (0.898–0.911)
ROC-AUC (sum of grade 2–4 probabilities)	0.940 (0.935–0.945)

4. Calibration maps and MC-Dropout passes

Table S13a. Calibration maps fitted on the calibration split of each fold, evaluated on the pooled out-of-fold predictions (95% CI, 500 clustered bootstrap resamples).

Map	ECE	Class-wise ECE	CW-ECE mild	CW-ECE severe	NLL	Brier	Macro-F1
Temperature scaling	0.0781 (0.0749–0.0854)	0.0770 (0.0752–0.0790)	0.0775 (0.0731–0.0818)	0.1022 (0.1001–0.1047)	0.6278 (0.6070–0.6467)	0.2827 (0.2734–0.2912)	0.7054 (0.6923–0.7177)
Vector scaling	0.0220 (0.0175–0.0275)	0.0088 (0.0081–0.0112)	0.0047 (0.0041–0.0091)	0.0038 (0.0031–0.0061)	0.3822 (0.3660–0.3968)	0.1931 (0.1849–0.2002)	0.7089 (0.6933–0.7250)
Dirichlet (ODIR)	0.0177 (0.0141–0.0237)	0.0078 (0.0074–0.0103)	0.0057 (0.0043–0.0100)	0.0033 (0.0025–0.0057)	0.3749 (0.3594–0.3889)	0.1924 (0.1843–0.1997)	0.7058 (0.6908–0.7218)
Isotonic	0.0085 (0.0064–0.0151)	0.0127 (0.0122–0.0151)	0.0164 (0.0134–0.0201)	0.0064 (0.0056–0.0090)	0.4243 (0.4050–0.4447)	0.1988 (0.1913–0.2056)	0.7258 (0.7106–0.7391)
Temperature-scaled MC-Dropout	0.0779 (0.0746–0.0853)	0.0778 (0.0759–0.0797)	0.0779 (0.0736–0.0822)	0.1032 (0.1011–0.1057)	0.6239 (0.6035–0.6424)	0.2827 (0.2734–0.2911)	0.7046 (0.6918–0.7171)

Table S13b. Paired differences from temperature scaling (95% CI).

Paired difference	ECE	Class-wise ECE	NLL
Dirichlet (ODIR) − temperature scaling	-0.0604 (-0.0693–-0.0542)	-0.0692 (-0.0706–-0.0658)	-0.2530 (-0.2654–-0.2400)
Isotonic − temperature scaling	-0.0697 (-0.0762–-0.0627)	-0.0643 (-0.0653–-0.0615)	-0.2036 (-0.2179–-0.1890)
Temperature-scaled MC-Dropout − temperature scaling	-0.0002 (-0.0037–0.0021)	0.0008 (0.0006–0.0010)	-0.0039 (-0.0047–-0.0031)
Vector scaling − temperature scaling	-0.0561 (-0.0648–-0.0505)	-0.0681 (-0.0697–-0.0650)	-0.2457 (-0.2585–-0.2326)

Table S14. Sensitivity to the number of MC-Dropout passes (mean ± SD over three random draws of M passes from 100 stored passes per image).

M	ECE	NLL	Macro-F1	MI error AUROC	Mean MI	Referable false-clear at τ = 0.80
5	0.0771 ± 0.0010	0.6259 ± 0.0010	0.7035 ± 0.0017	0.679 ± 0.0018	0.00516	1.45%
10	0.0791 ± 0.0001	0.6246 ± 0.0003	0.7040 ± 0.0005	0.721 ± 0.0041	0.00579	1.46%
20	0.0786 ± 0.0004	0.6239 ± 0.0004	0.7042 ± 0.0003	0.746 ± 0.0004	0.00612	1.48%
30	0.0780 ± 0.0006	0.6238 ± 0.0001	0.7048 ± 0.0007	0.758 ± 0.0015	0.00623	1.43%
50	0.0788 ± 0.0001	0.6235 ± 0.0001	0.7052 ± 0.0002	0.767 ± 0.0013	0.00633	1.45%

5. Training recipes, backbones and cost

Table S15a. Full comparison of training recipes and backbones on the same folds (C = temperature-scaled MC-Dropout; Δ = paired difference from the main model with 95% clustered bootstrap interval).

Model	Params (M)	Mean T	Mean epoch kept	Macro-F1	Bal. acc.	MCC	ECE A / B / C	CW-ECE (C)	CW-ECE severe	Max no-DR prob.	Δ macro-F1	Δ ECE	Δ CW-ECE
EffNet-B0, weighted CE (main)	4.34	0.678	12.0	0.705	0.769	0.738	0.186 / 0.078 / 0.078	0.078	0.103	0.895	+0.0000 [+0.0000, +0.0000]	+0.0000 [+0.0000, +0.0000]	+0.0000 [+0.0000, +0.0000]
EffNet-B0, plain CE	4.34	0.974	12.0	0.726	0.703	0.803	0.030 / 0.030 / 0.030	0.015	0.009	0.982	+0.0217 [+0.0087, +0.0351]	-0.0480 [-0.0581, -0.0422]	-0.0631 [-0.0649, -0.0602]
EffNet-B0, focal loss	4.34	0.731	12.0	0.726	0.704	0.796	0.079 / 0.018 / 0.018	0.014	0.007	0.996	+0.0214 [+0.0077, +0.0352]	-0.0594 [-0.0669, -0.0527]	-0.0638 [-0.0651, -0.0614]
EffNet-B0, class-balanced loss	4.34	0.822	12.0	0.734	0.755	0.785	0.050 / 0.022 / 0.023	0.032	0.041	0.959	+0.0299 [+0.0202, +0.0404]	-0.0550 [-0.0639, -0.0483]	-0.0456 [-0.0467, -0.0440]
EffNet-B0, sqrt oversampling	4.34	1.046	12.0	0.728	0.721	0.789	0.025 / 0.029 / 0.029	0.019	0.012	0.974	+0.0233 [+0.0116, +0.0356]	-0.0488 [-0.0580, -0.0420]	-0.0590 [-0.0601, -0.0564]
EffNet-B0, early stopping	4.34	0.690	15.2	0.719	0.743	0.760	0.196 / 0.084 / 0.084	0.069	0.098	0.864	+0.0147 [+0.0044, +0.0257]	+0.0063 [-0.0017, +0.0128]	-0.0089 [-0.0101, -0.0077]

ResNet-50	24.03	0.637	12.0	0.706	0.792	0.742	0.184 / 0.063 / 0.066	0.082	0.105	0.921	+0.0015 [-0.0092, +0.0115]	-0.0116 [-0.0184, -0.0045]	+0.0042 [+0.0027, +0.0056]
ViT-B/16	86.00	1.360	12.0	0.179	0.281	0.128	0.111 / 0.094 / 0.096	0.176	0.192	0.768	-0.5257 [-0.5397, -0.5114]	+0.0181 [+0.0094, +0.0256]	+0.0983 [+0.0963, +0.1006]

Table S15b. Training time on an NVIDIA GeForce RTX 5080 Laptop GPU (five folds).

Model	Total (h)	Per fold (min, mean ± SD)	Epochs run (mean)	Min per epoch	ms per image-epoch	Relative to main
EffNet-B0, weighted CE (main)	4.28	51.3 ± 3.1	12.0	4.28	29.9	×1.00
EffNet-B0, plain CE	4.48	53.8 ± 3.9	12.0	4.48	31.3	×1.05
EffNet-B0, focal loss	4.31	51.7 ± 1.2	12.0	4.31	30.1	×1.01
EffNet-B0, class-balanced loss	4.10	49.1 ± 3.8	12.0	4.10	28.6	×0.96
EffNet-B0, sqrt oversampling	4.62	55.5 ± 7.6	12.0	4.62	32.3	×1.08
EffNet-B0, early stopping	6.98	83.8 ± 11.8	18.8	4.44	34.5	×1.63
ResNet-50	6.41	76.9 ± 41.1	12.0	6.41	44.8	×1.50
ViT-B/16	4.01	48.1 ± 5.0	12.0	4.01	28.0	×0.94

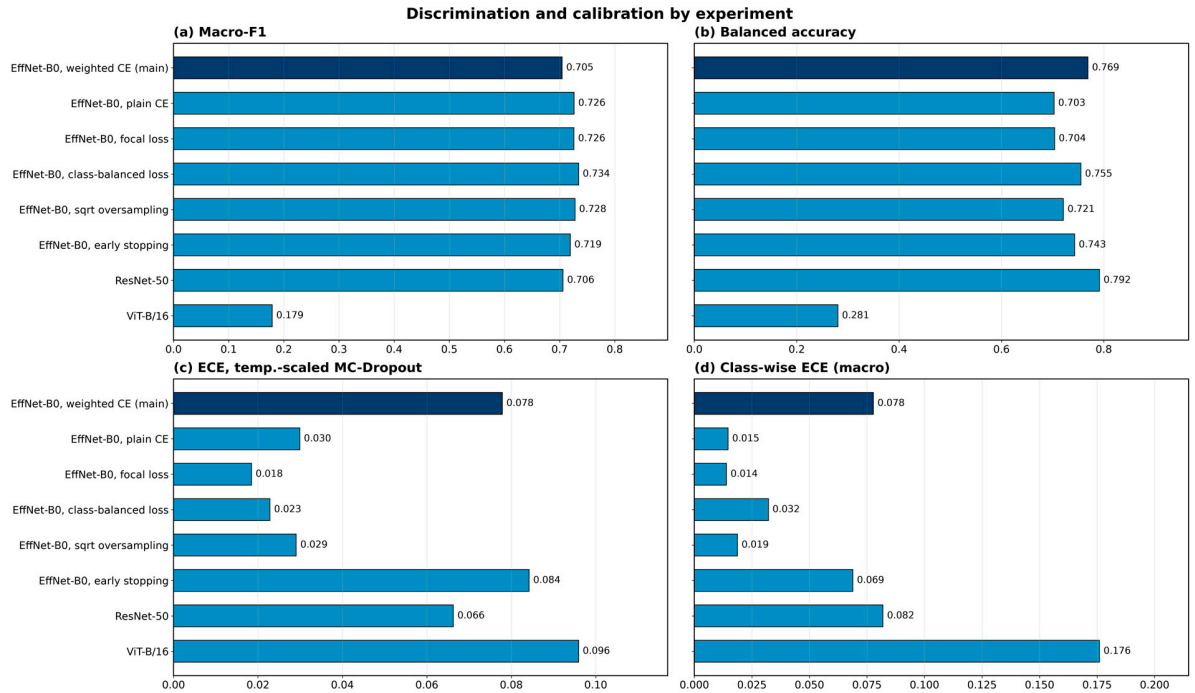


Figure S4. Macro-F1, balanced accuracy, ECE of variant C and macro class-wise ECE for each model.

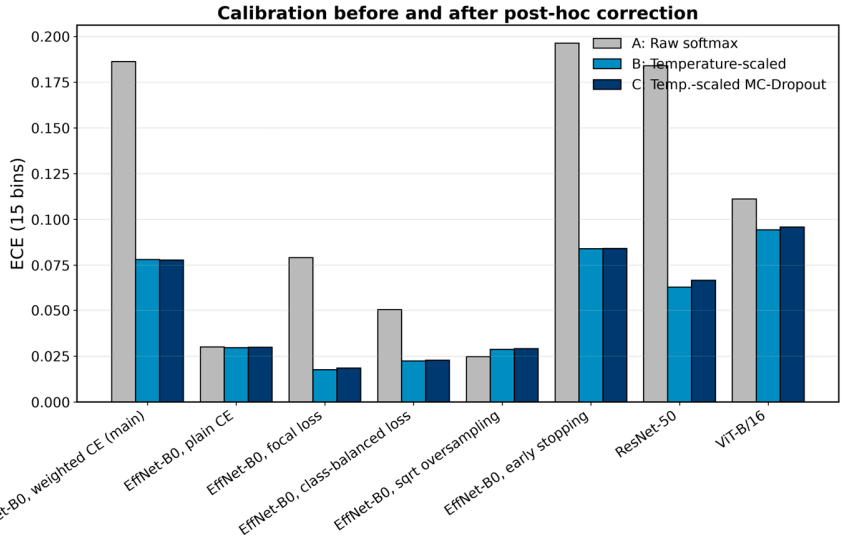


Figure S5. ECE of the raw output (A), after temperature scaling (B) and after temperature-scaled MC-Dropout (C) for each model.

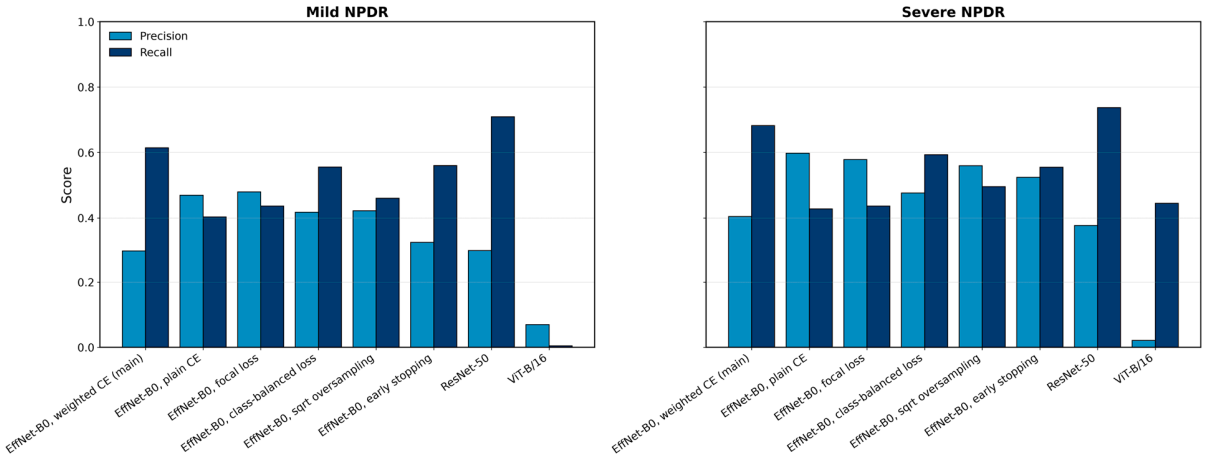


Figure S6. Precision and recall for mild and severe NPDR for each model.

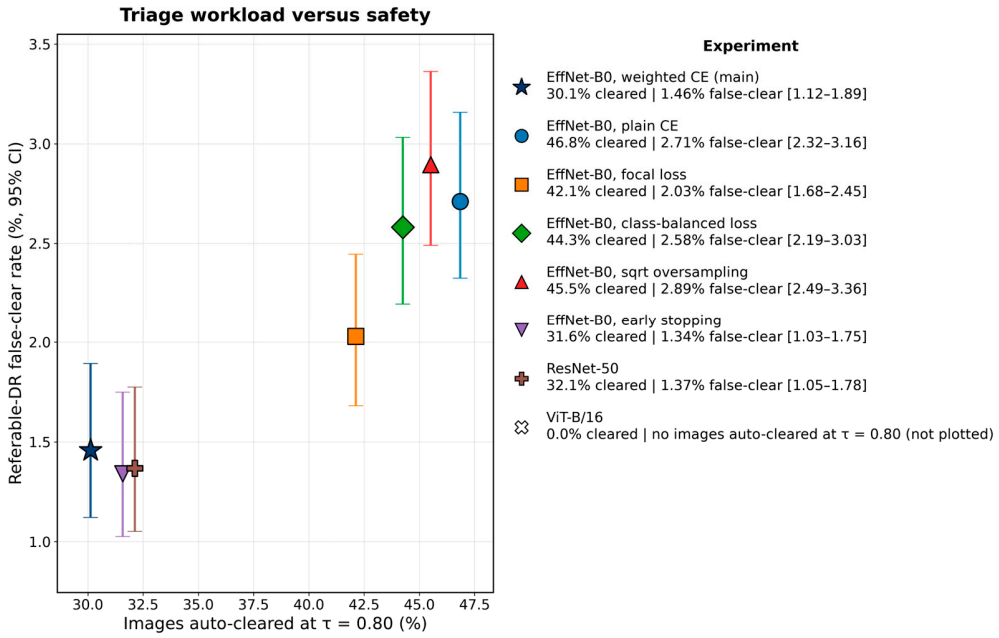


Figure S7. Auto-clear fraction and referable-DR false-clear rate (95% Wilson interval) at τ = 0.80 for each model. ViT-B/16 cleared no images.

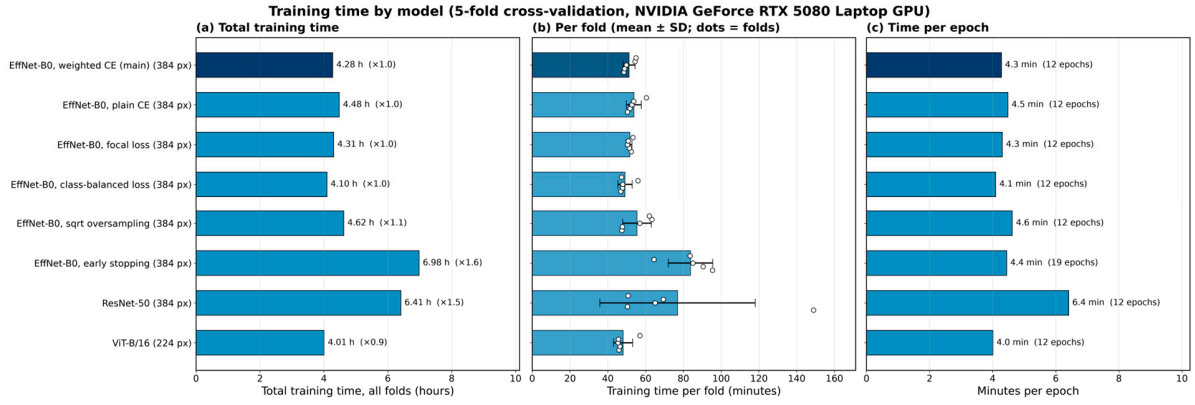


Figure S8. Training time by model: (a) total for five folds, (b) per fold with individual folds, (c) per epoch.

Table S16. Measured inference cost of the main model on an NVIDIA GeForce RTX 5080 Laptop GPU (15 warm-up and 50 timed runs; end-to-end rows include reading and preprocessing 100 real images).

Batch	Method	ms / image (mean)	ms / image (95th pct)	Images / s	Peak GPU memory (MB)
1	deterministic	8.54	11.08	117.1	1,003
1	mc dropout 20 full passes	107.19	124.62	9.3	1,003
1	mc dropout 20 backbone once	8.14	10.46	122.9	1,003
16	deterministic	1.00	1.20	999.3	1,467
16	mc dropout 20 full passes	19.54	19.65	51.2	1,467
16	mc dropout 20 backbone once	1.08	1.14	929.1	1,467

1	deterministic (end-to-end incl. loading and preprocessing)	32.56	124.09	30.7	
1	mc dropout 20 backbone once (end-to-end incl. loading and preprocessing)	35.92	133.58	27.8	

6. Explainability

Table S17. Grad-CAM statistics for the systematic sample of 110 test images, by true grade and correctness (each image explained by its own fold model; misclassified groups include the ten most confident errors).

True grade	Correct	Images	Area with CAM ≥ 0.5 (median)	Energy outside fundus (median)	Normalised entropy (median)	Peak inside fundus
Mild NPDR	no	10	0.028	0.072	0.899	100%
Mild NPDR	yes	10	0.027	0.148	0.941	90%
Moderate NPDR	no	18	0.039	0.058	0.932	94%
Moderate NPDR	yes	10	0.029	0.102	0.926	90%
No DR	no	10	0.025	0.159	0.918	100%
No DR	yes	10	0.058	0.118	0.935	100%
Proliferative DR	no	12	0.030	0.016	0.910	92%
Proliferative DR	yes	10	0.032	0.077	0.910	100%
Severe NPDR	no	10	0.023	0.018	0.887	100%
Severe NPDR	yes	10	0.042	0.029	0.934	100%

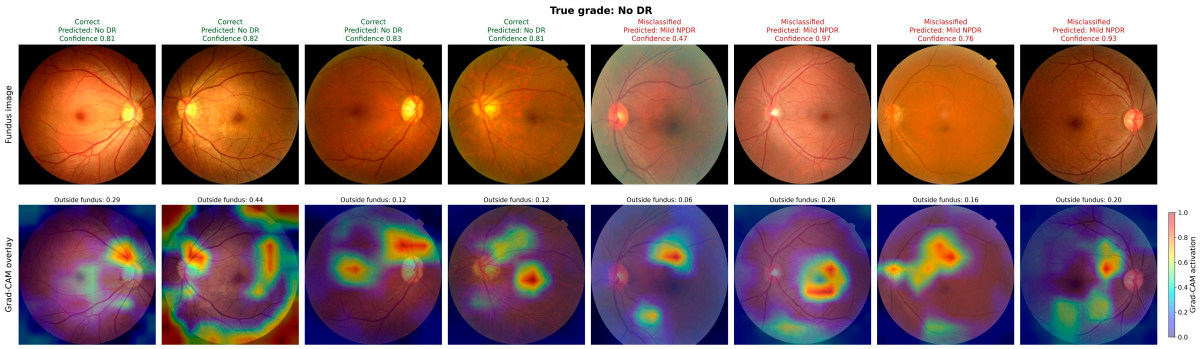


Figure S9a. Grad-CAM for true No DR test images: up to four correctly classified and four misclassified images (top: fundus; bottom: overlay; number above each overlay: share of activation energy outside the fundus).

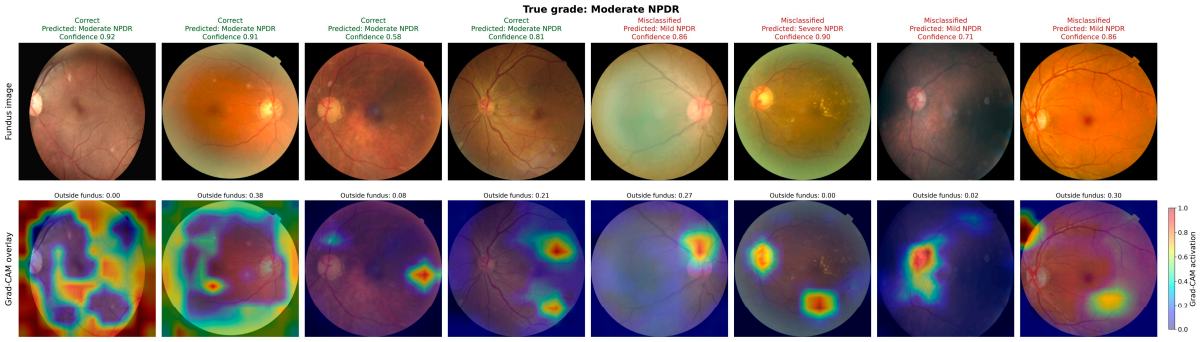


Figure S9b. Grad-CAM for true Moderate NPDR test images: up to four correctly classified and four misclassified images (top: fundus; bottom: overlay; number above each overlay: share of activation energy outside the fundus).

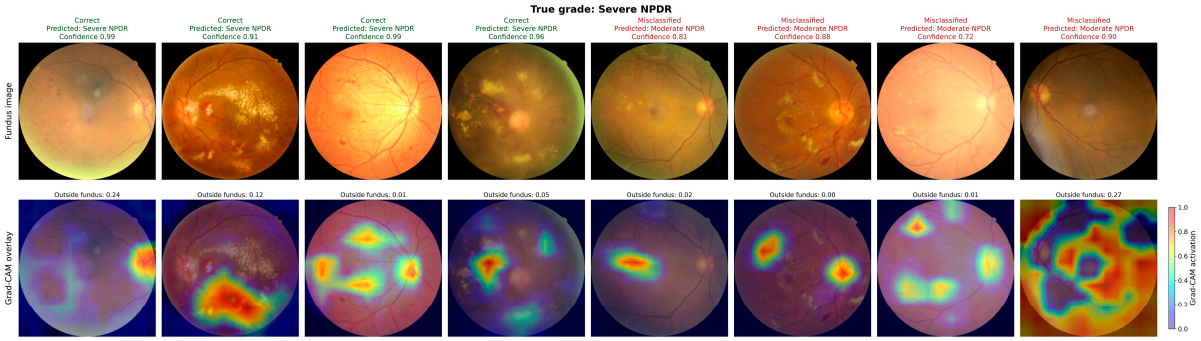


Figure S9c. Grad-CAM for true Severe NPDR test images: up to four correctly classified and four misclassified images (top: fundus; bottom: overlay; number above each overlay: share of activation energy outside the fundus).

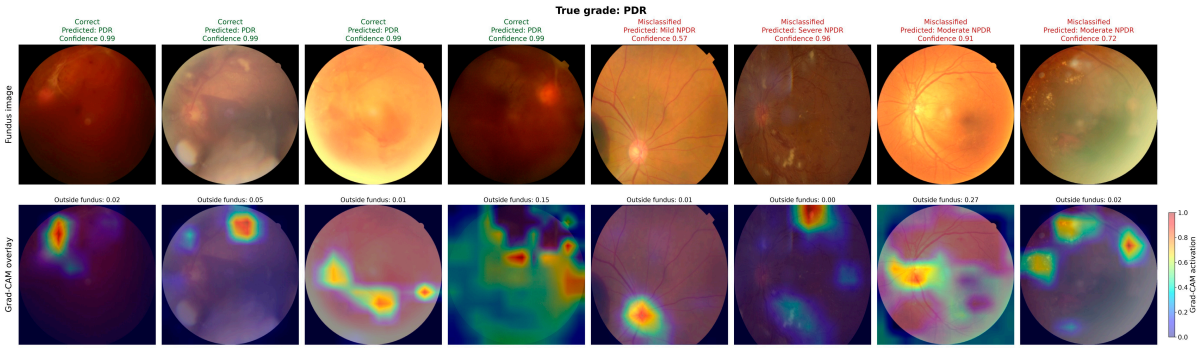


Figure S9d. Grad-CAM for true Proliferative DR test images: up to four correctly classified and four misclassified images (top: fundus; bottom: overlay; number above each overlay: share of activation energy outside the fundus).

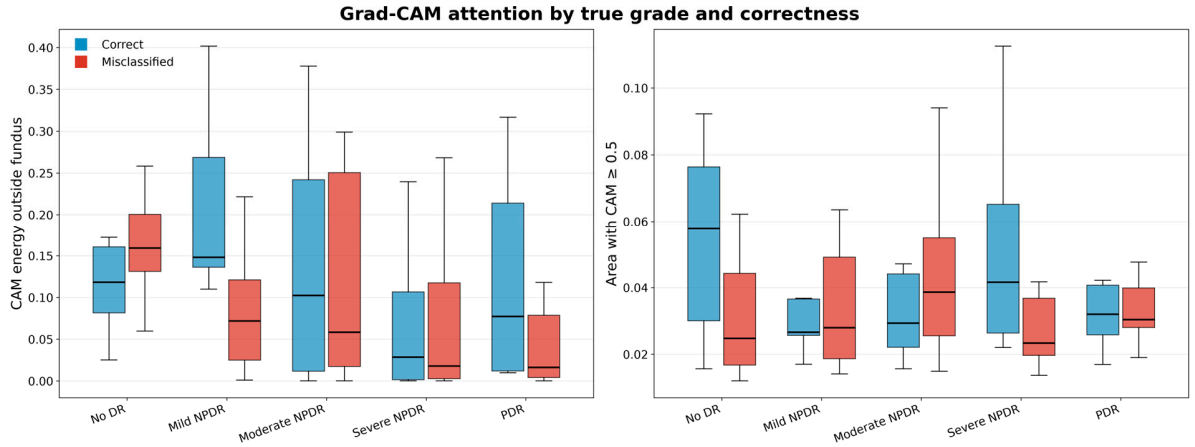


Figure S10. Grad-CAM activation energy outside the fundus (left) and area above 0.5 (right) by true grade and correctness.

Table S18. LIME explanation quality for ten images (most confident correct and most confident misclassified image per grade).

Measure	Mean (95% CI)	Median (IQR)	Correct / misclassified (mean)
Surrogate R^2	0.499 (0.401–0.590)	0.518 (0.420–0.596)	0.539 / 0.459
IoU between seeds	0.498 (0.400–0.591)	0.511 (0.381–0.601)	0.509 / 0.488
Dice between seeds	0.647 (0.562–0.729)	0.675 (0.545–0.748)	0.663 / 0.631
Spearman between seeds	0.605 (0.518–0.698)	0.579 (0.567–0.717)	0.614 / 0.596
IoU, 10% darker image	0.869 (0.824–0.917)	0.856 (0.844–0.881)	0.857 / 0.882
Dice, 10% darker image	0.928 (0.901–0.956)	0.922 (0.915–0.937)	0.921 / 0.936
Spearman, 10% darker image	0.971 (0.958–0.981)	0.975 (0.961–0.984)	0.978 / 0.963
Explained area (share of image)	0.221 (0.212–0.232)	0.218 (0.208–0.233)	0.221 / 0.221
Explained region outside fundus	0.030 (0.002–0.070)	0.000 (0.000–0.023)	0.009 / 0.052
Comprehensiveness (probability drop, explained removed)	0.767 (0.638–0.887)	0.824 (0.726–0.945)	0.678 / 0.855
Sufficiency (drop, only explained kept)	0.275 (0.067–0.523)	0.037 (0.004–0.593)	0.171 / 0.378
Probability drop, random superpixels removed	0.208 (0.051–0.407)	0.079 (0.002–0.378)	0.164 / 0.252
Comprehensiveness minus random	0.559 (0.402–0.708)	0.615 (0.349–0.763)	0.514 / 0.604
LIME–Grad-CAM IoU	0.165 (0.108–0.221)	0.187 (0.070–0.226)	0.156 / 0.174
LIME–Grad-CAM Dice	0.272 (0.179–0.357)	0.315 (0.129–0.368)	0.255 / 0.288
LIME–Grad-CAM Spearman	0.091 (-0.021–0.222)	0.037 (-0.068–0.279)	0.040 / 0.143

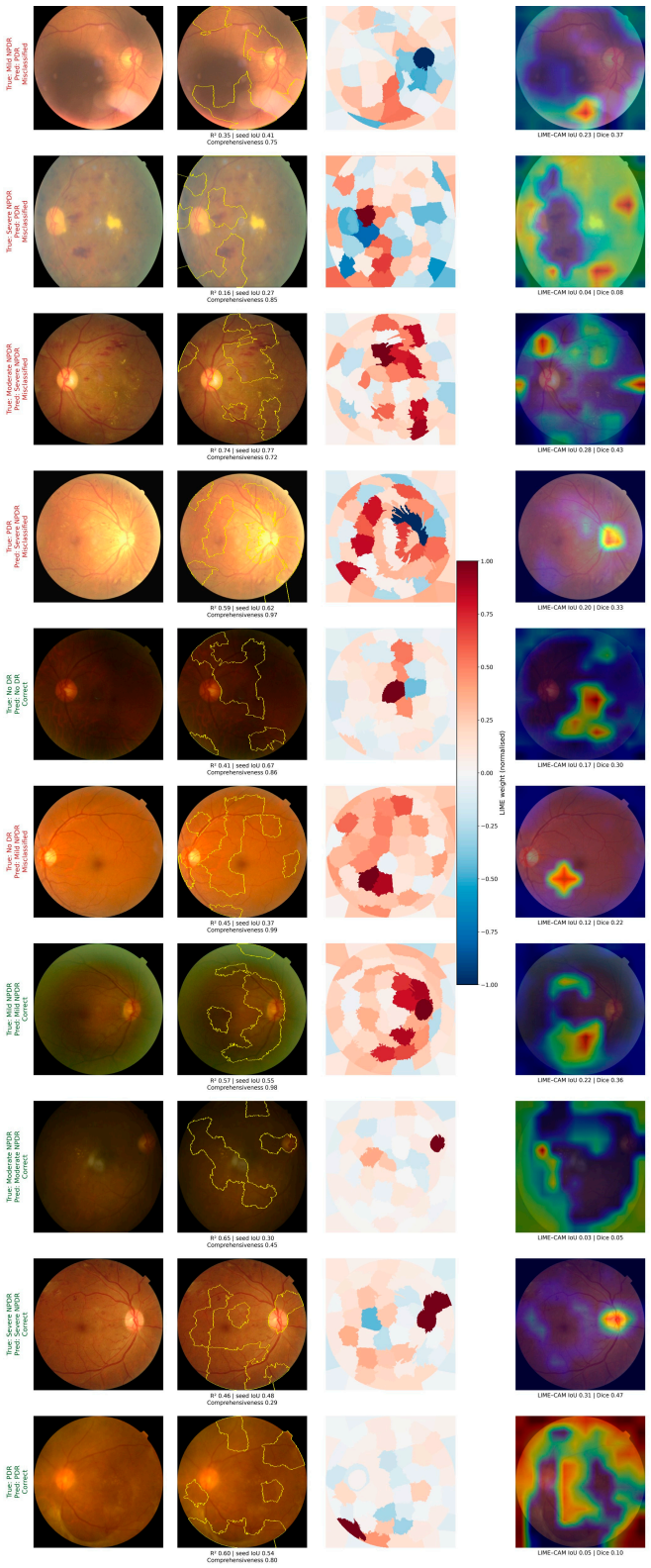


Figure S11. Individual LIME explanations: fundus image, LIME region (12 top superpixels), LIME attribution map and Grad-CAM for the ten explained images; true and predicted grades on the left.

7. Data and code

Data S1: all_oof_predictions_r0.csv — out-of-fold probabilities of all variants and calibration maps, uncertainty scores, fold, temperature and near-duplicate group for each of the 12,522 images.

Data S2: splits_group_repeat0.csv — fold and partition (inner training, calibration, outer test) of every image, with its near-duplicate group.

Data S3: all_errors_sorted_by_confidence.csv — the 2,127 misclassified images with confidence and uncertainty, sorted by confidence.

Code S1: DDR_Revision_Training.ipynb (training and evaluation notebook) and preprocessing.py (stand-alone preprocessing used for training and inference).