

Do Age and Sex Help Deep Fundus Disease Classifiers? A Controlled Analysis of Metadata Fusion and Attention on ODIR-5K

Safeer Gohar^{1,2} , Touqir Gohar², Sardar Ahamd Ali³

1. College of Computer Science and Cybersecurity, Chengdu University of Technology, Chengdu 610059, China

2. Department of Computer Engineering, University of Engineering and Technology, Peshawar, Pakistan

3. Department of Computer Sciences and Engineering, AIR University, Islamabad, Pakistan

4. E-mail any correspondence to: Safeer Gohar (safeergohar77@gmail.com)

How to cite: Gohar S, Gohar T, Ali SA. Do Age and Sex Help Deep Fundus Disease Classifiers? A Controlled Analysis of Metadata Fusion and Attention on ODIR-5K. *Innov Res Appl Biol Chem Sci (IRABCS)*. 2025;3(2):65-71. Available from: <https://irjpl.org/irabcs/article/view/278>. DOI: <https://doi.org/10.62497/irabcs.278>.

Abstract

Patient age and sex are mostly fused into medical image classifiers, and mostly small increases occur in accuracy. Such gains are rarely compared with what the metadata can predict alone, or tested for whether the network actually uses the metadata. The study was conducted using the ODIR-5K dataset (3,500 patients and eight labels), employing a patient-level split and three models trained under identical conditions with three random seeds: metadata only, a dual-eye ResNet-50 on fundus images only, and the same network fused with age and sex embeddings. Age and sex alone reached a macro-AUC of 0.587; the image-only model reached 0.863 and the fused model 0.861 (difference -0.002 , 95% CI -0.012 to 0.007), with no label improving after Holm correction. When test metadata were shuffled, macro-AUC fell by

only 0.0015, and sweeping a patient's age from under 30 to over 80 left the predicted cataract probability almost unchanged (0.064 to 0.065), although cataract prevalence rises from 2% to 37% over that range. Stacking age and sex on the image model's outputs gave no gain either. Quantified Grad-CAM showed that fusion alters attention no more than retraining with a different random seed. For this task, age and sex were considered redundant when image-based evidence was available; therefore, it is recommended that metadata-fusion studies report a metadata-only control and conduct a metadata-reliance test.

Keywords: fundus imaging, multi-label classification, metadata fusion, explainable AI, Grad-CAM, ODIR-5K

Introduction

Deep learning on color fundus photographs now detects diabetic retinopathy and other ocular diseases at specialist level [1], and multi-disease benchmarks such as ODIR-5K have made joint screening for several conditions a standard task [2]. A recurring design choice in medical imaging is to fuse the image with structured patient information. Age and sex are the most common additions because they are nearly always recorded, and reviews of image-record fusion report consistent, if modest, improvements [3]. In dermatology, the inclusion of clinical data alongside lesion images has demonstrated enhanced categorization [4, 5].

Such studies usually miss two checks. The first is a metadata-only control: without knowing how well age

and sex predict each disease on their own, a fused model's gain cannot be attributed to the image, to the metadata, or to their combination. The second is a dependency assessment: a fused network may improve, deteriorate, or remain unchanged for reasons unrelated to its utilization of metadata. A related issue is the analysis of explainability, as saliency maps vary for otherwise identical models, and the discrepancies between two maps lack utility without considering that fluctuation [6].

Both checks were applied to age–sex fusion for multi-label fundus classification using the ODIR-5K dataset. Three models were compared under identical training conditions, and four research questions were investigated. (E1) Does fusion improve per-disease AUC

once the metadata-only control is known? (E2) Where in the difficulty spectrum does any gain arise? (E3) Does fusion change *where* the network looks? (E4) Does the fused network actually use age and sex? Three control studies were incorporated to differentiate architectural factors from informational ones. Our findings indicate a robust null hypothesis: age and gender contribute no more quantifiable data to a dual-eye vision model within this dataset, and the integrated network effectively disregards these factors.

Fundus classification

Colour fundus photographs (CFPs) are utilised to identify significant diabetic retinopathy [1] and other concomitant ocular conditions [7] using deep learning networks. In ODIR-5K, the objective is defined as multi-label identification at the patient level utilizing data from both eyes [2]. The majority of delineated models rely solely on images. As the labels characterize the patient, managing patient-level data is crucial; segregating images instead of patients would enable the two eyes of a single individual to be included in both the training and testing datasets.

Metadata fusion

Systematic reviews of image–record fusion distinguish early, joint, and late fusion and report that adding structured data usually helps, while noting that simple baselines are often missing [3]. The classification of lesions in dermatology is refined through patient and lesion metadata, including age, sex, and body site, although this material includes details not visible in the image. Fundus images encapsulate demographic data: dedicated networks proficiently determine age and sex from retinal photographs [8, 9], while colonics expand this analysis to systemic health conditions [10]. Therefore, it is an empirical matter to ascertain if a disease classifier enhances its performance when provided with age and sex information. This research analyses it explicitly.

Explainability

Grad-CAM is the default tool for showing where fundus models look. Saliency maps could remain unresponsive to the model [6] and inconsistent for localization in medical images [11]. A network can achieve elevated accuracy by employing shortcuts instead of relying on pathologies [12]. Thus, statistically assessing attentiveness and contrasting it with the variability among random seeds.

Materials and Methods

Dataset and split

ODIR-5K comprises left and right fundus images from 3,500 individuals, accompanied by patient-level classifications across eight categories: normal (N), diabetic retinopathy (D), glaucoma (G), cataract (C), age-related macular degeneration (A), hypertensive retinopathy (H), pathological myopia (M), and miscellaneous abnormalities (O). Patient age ranges from 1 to 91 years, with a median of 59, and sex is recorded, showing 46.1% female. Since the labels are positioned at the patient level rather than at eye level,

the dataset was divided at the patient level into training (2,449), validation (351), and test (700) subsets, ensuring stratification based on the label combinations. No patient or image appears in more than one split, and age and sex distributions do not differ between training and test (Kolmogorov–Smirnov $p = 0.89$; $\chi^2 p = 0.21$). Before any model was trained, the expected size of the fusion gain was recorded for each label (large for C and A, moderate for G and H, small for D and N, small or negative for M), together with a timestamp and hash.

The annotations flag lens dust for 265 patients and low image quality for 19; all images were kept, since such artefacts are part of the screening task and affect all three models equally. **Table 1** lists label prevalence by age decade. Cataract, glaucoma and AMD become more common with age, whereas pathological myopia is concentrated in the youngest group. These are exactly the patterns a metadata branch could exploit.

Table 1: Label prevalence (%) by age decade, all patients

Age	n	N	D	G	C	A	H	M	O
<30	51	31.4	17.6	2.0	2.0	2.0	0.0	27.5	25.5
30–39	153	39.2	35.3	6.5	3.3	2.6	2.6	3.9	19.6
40–49	537	37.4	36.7	3.7	1.7	3.0	3.4	2.2	23.5
50–59	110	34.9	37.2	3.8	3.5	4.6	3.4	4.2	25.8
60–69	116	30.1	31.5	6.9	5.7	5.4	3.6	4.5	32.2
70–79	40	27.0	20.1	12.5	15.0	5.2	0.2	9.8	32.4
80+	81	21.0	8.6	13.6	37.0	9.9	0.0	3.7	21.0
All	3500	32.6	32.2	6.1	6.1	4.7	2.9	5.0	28.0

Models

All three models predict the eight labels jointly. The metadata-only model is a per-label logistic regression on one-hot age decade (seven bins) and sex, i.e. exactly the information given to the fused network; gradient boosting on continuous age served as a sensitivity check. The image-only model is a dual-eye network: a shared ImageNet-pretrained ResNet-50 [13,14] encodes each eye, the two pooled 2,048-d vectors are concatenated and projected to 256 dimensions, and a linear head outputs eight logits. The fused model is identical except that 16-d learned embeddings of age decade and sex are concatenated to the 256-d image representation before the head (**Figure 1**). This simple late fusion was used so that results speak about the metadata rather than about a fusion module.

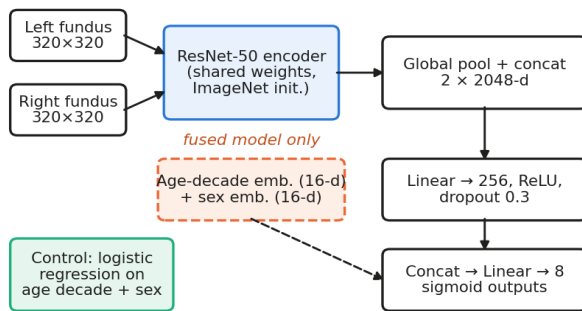


Figure 1: The three models. Image-only and fused networks are identical except for the dashed metadata branch; the metadata-only control receives the same age and sex encoding.

Training

Images were cropped to the retinal field of view, padded to square and resized to 320×320 pixels. Image-only and fused models were trained with identical settings: binary cross-entropy, AdamW (learning rate 10^{-4} , weight decay 10^{-4}) [15] with a one-cycle schedule, batches of 12 patients, mixed precision, light augmentation (random resized crop, flip, $\pm 15^\circ$ rotation, brightness/contrast jitter), up to 30 epochs, and early stopping on validation macro-AUC (patience 7). Each model was trained with three seeds (0, 1, 2) on a single NVIDIA RTX 5060 laptop GPU.

The code is implemented in PyTorch with the timm model library and is released with a single script that reproduces every table and figure from the raw dataset. Each of the six image models was trained in 1.3–1.9 hours. ODIR-5K is a public, de-identified dataset, and no additional ethical approval was required.

Analyses

E1 reports test AUC per label for each model. Primary estimates use the average of the three seeds' predicted probabilities; per-seed means and standard deviations are also reported. 95% confidence intervals come from 2,000 stratified bootstrap resamples [16], paired comparisons use DeLong's test [17,18], and p -values are Holm-corrected across the eight labels [19].

E2 divides test patients into thirds by the per-patient loss of image-only seed 0 and compares the remaining seeds of the fused and image-only models within each third; defining difficulty with a model that is not part of

the comparison avoids a spurious regression-to-the-mean gain.

E3 computes Grad-CAM [20] maps for both eyes of every test patient and every label that patient has, and summarizes each map by the fraction of attention inside the retina, normalized spatial entropy, and the share of attention in the hottest 5% of pixels, comparing models with paired Wilcoxon tests.

E4 keeps each test image fixed and alters only the metadata given to the fused models: age and sex shuffled across patients (20 permutations), sex flipped, and age swept through every decade.

Three controls address alternative explanations.

(i) *Stacking*: a logistic regression per label on the image-only logit, with and without age and sex, fitted on the validation set and also cross-fitted (5-fold) on validation plus test, tests whether the metadata carry information beyond the image model regardless of how fusion is built.

(ii) *Probing*: ridge and logistic regressions on the 256-d image representation test how much age and sex the image features encode.

(iii) *Grad-CAM seed control*: map correlations between image-only models trained with different seeds give the reference variation against which image-versus-fused differences are judged.

Results

Fusion does not improve classification (E1)

Age and sex alone were weak predictors (macro-AUC 0.587, 95% CI 0.556–0.615; gradient boosting 0.594). Their best label was cataract (0.732), and for AMD they were at chance (0.493). The image-only model reached 0.863 (0.845–0.881) and the fused model 0.861 (0.843–0.878), a difference of -0.002 (-0.012 to 0.007) (Table 2, Figure 2). Per-seed results agreed (0.845 ± 0.014 vs. 0.843 ± 0.006). No label changed significantly after correction; the largest point estimates were $+0.027$ for hypertensive retinopathy (CI -0.019 to 0.066) and -0.018 for normal (-0.037 to 0.000). The pre-registered ordering did not hold (Spearman $\rho = 0.09$, $p = 0.84$), unsurprisingly, since every gain was close to zero.

Table 2: Test AUC per label (seed-averaged; 700 patients)

Label	n+	Meta	Image	Fused	Δ fused – image [95% CI]
N Normal	228	0.546	0.781	0.763	-0.018 [-0.037 , 0.000]
D Diabetic ret.	227	0.559	0.812	0.805	-0.007 [-0.021 , 0.007]
G Glaucoma	46	0.625	0.939	0.927	-0.012 [-0.029 , 0.006]
C Cataract	42	0.732	0.972	0.985	$+0.012$ [-0.002 , 0.033]
A AMD	33	0.493	0.882	0.867	-0.015 [-0.044 , 0.013]
H Hypertensive	18	0.562	0.845	0.872	$+0.027$ [-0.019 , 0.066]
M Path. myopia	35	0.637	0.980	0.981	$+0.000$ [-0.031 , 0.032]
O Other	195	0.539	0.696	0.689	-0.007 [-0.030 , 0.017]
Macro		0.587	0.863	0.861	-0.002 [-0.012 , 0.007]

No difference is significant after Holm correction (smallest adjusted $p = 0.47$).

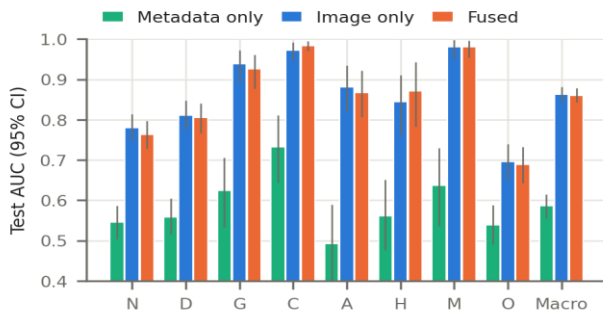


Figure 2: Test AUC with 95% bootstrap CIs for the metadata-only, image-only and fused models. Adding age and sex leaves every label within its confidence interval.

The fused network ignores age and sex (E4)

Shuffling age and sex across test patients reduced the fused model's macro-AUC by only 0.0015 ± 0.0008 (age alone 0.0010 , sex alone 0.0007 ; Figure 3), and flipping sex changed predicted probabilities by 0.001 – 0.019 on average.

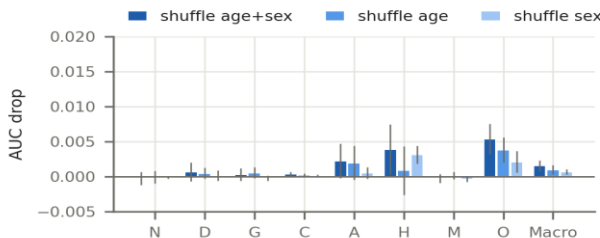


Figure 3: E4 drop in the fused model's test AUC when age and sex, age only, or sex only are shuffled across patients (mean $\pm$ SD over 20 permutations). Every drop is below 0.006 .

The age sweep makes the point most directly (Figure 4). Holding the image fixed and moving the recorded age from under 30 to over 80 changed the mean predicted cataract probability from 0.064 to 0.065 , whereas the observed prevalence of cataract across those groups rises from 2.0% to 37.0% . Glaucoma (0.057 to 0.060 against 2.0% to 13.6%) and pathological myopia (flat at 0.049 , although 27.5% of patients under 30 have it) behave the same way. The network has learned an almost constant contribution from the metadata branch.

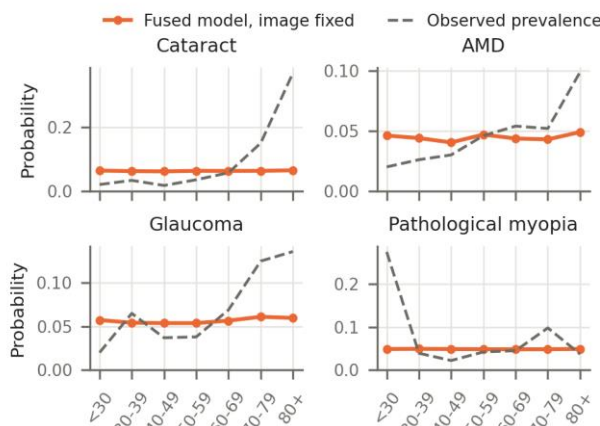


Figure 4: Age sweep. Mean predicted probability of

the fused model when the same test images are paired with each age decade (solid), against the observed prevalence in that decade (dashed).

The null does not depend on the fusion design (controls)

If the network had failed to exploit useful metadata, a simpler combination should recover the gain. It did not (Table 3). Cross-fitted stacking of age and sex on the image-only logits changed AUC by at most 0.013 for any label, and none was significant (all Holm $p > 0.95$). With a single validation-set fit, AUC even fell for AMD (-0.027), reflecting the cost of estimating twelve extra parameters on 351 patients. Probing showed that the image features encode age only weakly (correlation 0.34 – 0.35 , mean absolute error 8.8 years against 9.6 years for predicting the training mean) and sex barely above chance (AUC 0.57 – 0.58). Features from the ImageNet backbone without fine-tuning performed similarly ($r = 0.39$, sex AUC 0.56). The metadata are therefore not redundant because the network *re-derives* age and sex from the image. They are redundant because the disease itself is visible: once the lesion evidence is in the image, knowing the patient's age adds little to the ranking of cases.

Table 3: Metadata reliance and control analyses

Analysis	Result
Shuffle age + sex (fused, macro-AUC drop)	0.0015 ± 0.0008
Age sweep, cataract: Δp (80+ vs. <30)	$+0.001$ (prevalence $+35$ pts)
Stacking, cross-fit: largest $ \Delta \text{AUC} $	0.013 (all Holm $p > 0.95$)
Age probe r / MAE: image fused	$0.34 / 8.8 \text{ y} \mid 0.35 / 8.8 \text{ y}$
Age probe: ImageNet features mean	$0.39 / 8.7 \text{ y} \mid - / 9.6 \text{ y}$
Sex probe AUC: image fused ImageNet	$0.58 \mid 0.57 \mid 0.56$
Grad-CAM r : image vs. fused (same seed)	0.574
Grad-CAM r : image vs. image (other seed)	0.569 ($p = 0.84$)

Probe values are medians over three seeds; one image-only seed gave an unstable age probe ($r = 0.06$). Grad-CAM r : median pixel correlation over 4,800+ maps per pair type; p from Mann–Whitney test.

Fusion does not change where the network looks (E3)

Across 1,645 paired maps, attention stayed on the retina for both models (median in-retina fraction 0.991 vs. 0.989). Spatial entropy (0.861 vs. 0.862 , $p = 0.28$) and top-5% concentration (0.500 vs. 0.491 , $p = 0.35$) did not differ overall. Some per-label differences reached significance at this sample size, but all were small (the largest was a 0.020 lower in-retina fraction for glaucoma). The seed control shows how small they are: maps of the image-only and fused models trained with the same seed correlated as strongly (median 0.574) as maps of two image-only models trained with different seeds (0.569 ; $p = 0.84$). Fusion changes attention by no

more than retraining does, and in the examples of **Figure 5** both models highlight the same lesions.

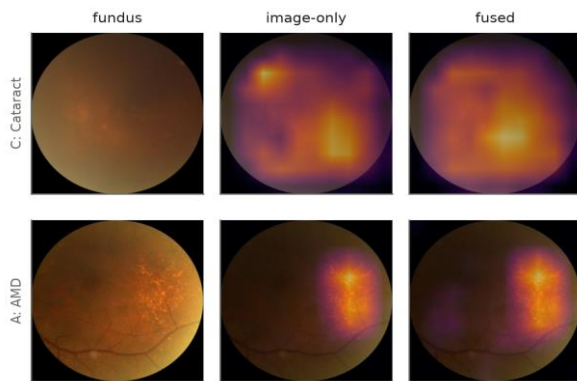


Figure 5: Grad-CAM for a cataract and an AMD test case. Image-only and fused models (same seed) attend to the same retinal regions.

Where the residual effect lies (E2) and subgroups

Within the hardest third of cases, fusion reduced the per-patient loss slightly (0.024, CI 0.003–0.046), while it increased the loss slightly among easy and intermediate cases (–0.012 and –0.011). The net effect cancels, consistent with E1 (**Figure 6**). Fusion did not significantly change performance gaps between subgroups. The male–female gap in macro-AUC moved from 0.023 to 0.008 (change –0.015, CI –0.034 to 0.004), and the gap between patients aged 65 or more and those under 50 moved from –0.033 to –0.028 (change 0.005, CI –0.019 to 0.027).

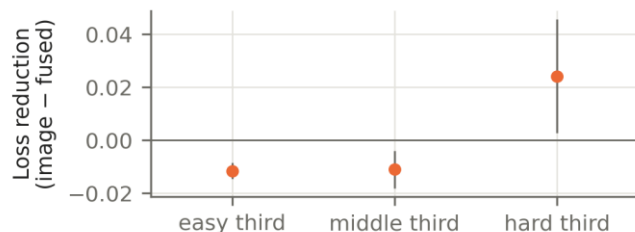


Figure 6: E2: mean reduction in per-patient loss from fusion within thirds of difficulty (defined by an independent image-only seed), pooled over labels, with 95% bootstrap CIs.

Discussion

The central finding is simple. On ODIR-5K, a carefully trained dual-eye image model leaves nothing for age and sex to add, and the fused network responds by ignoring them. Three pieces of evidence rule out the obvious alternative explanations. The metadata-only control shows that age and sex are weak predictors in the first place. The stacking control shows that the null does not depend on how fusion is built. The probe shows that the image features are not simply re-deriving age and sex. What remains is that, for diseases defined by visible fundus findings, the lesion carries the diagnosis, and a demographic prior changes the ranking of cases very little. This agrees with the limited cataract gain: although cataract depends strongly on age, the image-only model already reaches 0.972 AUC for it.

A short calculation shows why a strong demographic prior can leave the ranking almost unchanged. If image

and demographics were conditionally independent given the disease, Bayes' rule would give $\text{logit } P(y|x,a) = \text{logit } P(y|x) + [\text{logit } P(y|a) - \text{logit } P(y)]$. The metadata could then only shift each patient's log-odds by the ratio of age-specific to overall prevalence. For cataract, this shift differs by about 3.4 log-odds units between the youngest and oldest decades (prevalence 2% vs. 37%). With an image-only AUC of 0.972, however, 97% of positive–negative pairs are already correctly ordered, so only the few pairs whose image scores are close and whose ages differ can be reordered. Our stacking control fits exactly this additive model and finds no gain, and E2 locates the small residual benefit in the hardest cases, where such close pairs occur.

The practical implication concerns reporting. A fused model that is marginally better or worse than its image-only counterpart is uninterpretable without the metadata-only control, and a claimed benefit of fusion should be supported by a reliance test such as metadata shuffling or an attribute sweep, which here cost minutes of computation and required no retraining. The same holds for explainability. Our Grad-CAM metrics detected statistically significant differences between models, yet the seed control showed that these are no larger than the differences between two runs of the same model. Saliency comparisons should therefore be reported against a seed baseline, in line with earlier warnings about the reliability of saliency methods [6]. Grad-CAM also offers an additional check against shortcut learning [12], since attention remained on the retina and on the lesions.

Limitations

A single dataset was used whose age information is coarse and whose labels are patient-level, and a single backbone and late-fusion design; the stacking control mitigates, but does not remove, the design dependence. Datasets in which the target disease is subtle or invisible on the fundus, or in which metadata carry systemic risk, may benefit from fusion. Oculomics studies show that retinal images can predict systemic factors when models are trained for that purpose [8,10]. With 18 to 46 positive test cases, the confidence intervals for the rare labels (H, A, M, C, G) remain wide.

Future work should repeat the three checks on datasets with richer metadata, such as blood pressure, diabetes duration or HbA1c, and on tasks in which metadata carry information that the image cannot show, for example progression or systemic-risk prediction. It should also compare fusion at earlier network stages. The reliance test extends directly to any structured input, and the seed-referenced saliency comparison applies to any explanation method.

Conclusion

Adding patient age and sex to a dual-eye ResNet-50 on ODIR-5K did not improve any disease label, did not change where the network looks beyond seed-level variation, and yielded a network that almost ignores the metadata it is given. These conclusions were supported by a metadata-only control, a metadata-reliance test,

and seed-referenced explainability, all three of which are recommended as routine checks in metadata-fusion studies.

Authorship Contribution Statement

Safeer Gohar: Conceptualization; Methodology; Software; Investigation; Formal analysis; Data curation; Visualization; Writing – original draft; Writing – review & editing.

Touqir Gohar: Methodology; Validation; Formal analysis; Investigation; Writing – review & editing.

Sardar Ahmad Ali: Conceptualization; Methodology; Supervision; Validation; Writing – review & editing.

All authors contributed substantially to the study, participated in drafting or critically revising the manuscript, approved the final version for publication, and agreed to be accountable for all aspects of the work.

Availability of data and materials

The data in the current study is available from the corresponding author on reasonable request.

Conflict of interest

The authors reported no conflict of interest.

References

- [1] Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402–2410. doi:10.1001/jama.2016.17216
- [2] 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. Vol. 12614. Springer; 2021. p. 177–193. doi:10.1007/978-3-030-71058-3_11
- [3] Huang SC, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ Digit Med*. 2020;3:136. doi:10.1038/s41746-020-00341-z
- [4] Pacheco AGC, Krohling RA. The impact of patient clinical information on automated skin cancer detection. *Comput Biol Med*. 2020;116:103545. doi:10.1016/j.compbimed.2019.103545
- [5] Kawahara J, Daneshvar S, Argenziano G, Hamarneh G. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE J Biomed Health Inform*. 2019;23(2):538–546. doi:10.1109/JBHI.2018.2824327
- [6] Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. In: *Advances in Neural Information Processing Systems (NeurIPS)*; 2018. p. 9505–9515. arXiv:1810.03292. doi:10.48550/arXiv.1810.03292
- [7] Ting DSW, Cheung CYL, Lim G, Tan GSW, Quang ND, Gan A, 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(22):2211–2223. doi:10.1001/jama.2017.18152
- [8] Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng*. 2018;2(3):158–164. doi:10.1038/s41551-018-0195-0
- [9] Korot E, Pontikos N, Liu X, Wagner SK, Faes L, Huemer J, et al. Predicting sex from retinal fundus photographs using automated deep learning. *Sci Rep*. 2021;11(1):10286. doi:10.1038/s41598-021-89743-x
- [10] Wagner SK, Fu DJ, Faes L, Liu X, Huemer J, Khalid H, et al. Insights into systemic disease through retinal imaging-based oculomics. *Transl Vis Sci Technol*. 2020;9(2):6. doi:10.1167/tvst.9.2.6
- [11] Arun N, Gaw N, Singh P, Chang K, Aggarwal M, Chen B, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiol Artif Intell*. 2021;3(6):e200267. doi:10.1148/ryai.2021200267
- [12] Geirhos R, Jacobsen JH, Michaelis C, Zemel R, Brendel W, Bethge M, et al. Shortcut learning in deep neural networks. *Nat Mach Intell*. 2020;2:665–673. doi:10.1038/s42256-020-00257-z
- [13] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2016. p. 770–778. doi:10.1109/CVPR.2016.90

- [14] Deng J, Dong W, Socher R, Li LJ, 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); 2009. p. 248–255. doi:10.1109/CVPR.2009.5206848
- [15] Loshchilov I, Hutter F. Decoupled weight decay regularization. In: Proceedings of the International Conference on Learning Representations (ICLR); 2019. arXiv:1711.05101. doi:10.48550/arXiv.1711.05101
- [16] Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman & Hall; 1993. doi:10.1201/9780429246593
- [17] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–845. doi:10.2307/2531595
- [18] Sun X, Xu W. Fast implementation of DeLong’s algorithm for comparing the areas under correlated receiver operating characteristic curves. IEEE Signal Process Lett. 2014;21(11):1389–1393. doi:10.1109/LSP.2014.2337313
- [19] Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6(2):65–70. doi:10.2307/4615733
- [20] Selvaraju RR, 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); 2017. p. 618–626. doi:10.1109/ICCV.2017.74

Disclaimer: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.