Beyond External Validation: Evaluating the Deployment Readiness of Medical Imaging AI in an African Clinical Context through Diabetic Retinopathy Screening

Kerol Djoumessi[1(✉)][0009−0004−1548−9758], Ciku Wanjiku Mathenge[2], and Philipp Berens[1,3][0000−0002−0199−4727]

[1] Hertie Institute for AI in Brain Health, University of Tübingen, Germany
kerol.djoumessi-donteu@uni-tuebingen.de
[2] Rwanda International Institute of Ophthalmology
[3] Tübingen AI Center, University of Tübingen, Germany

Abstract. Deep learning models for medical imaging are increasingly assessed across multiple external datasets before clinical deployment. However, strong external validation does not necessarily demonstrate deployment readiness, particularly in underrepresented healthcare settings. Using diabetic retinopathy (DR) screening as a case study, we evaluate two publicly available models—an interpretable sparse BagNet and a conventional ResNet-50—originally trained on the Kaggle dataset and externally validated on ten publicly available retinal imaging datasets. Without retraining or adaptation, both models were assessed on a private retinal fundus dataset from the Nakuru Eye Disease Study in Kenya using the original evaluation protocol. After reproducing the published external validation results, evaluation on the Kenyan cohort revealed substantial reduced sensitivity while specificity remained high. The reduced sensitivity was entirely driven by missed Grade 1 DR cases, with no false negatives for Grades 2–4. Threshold optimization partially improved sensitivity but did not recover previous external validation performance, while qualitative analysis of the interpretable model provided complementary evidence for local model auditing. These findings demonstrate how local cross-context evaluation, operating-threshold assessment, grade-specific error analysis, and qualitative auditing can complement conventional external validation and provide context-specific evidence to inform deployment decisions in underrepresented healthcare settings.

Keywords: Medical imaging AI · Deployment readiness · External validation · Diabetic Retinopathy Detection · African healthcare.

1 Introduction

Artificial intelligence (AI) has achieved remarkable progress in medical image analysis, enabling automated disease detection and diagnosis across diverse clinical applications, including radiology and ophthalmology [12, 18]. As these systems increasingly demonstrate expert-level performance on benchmark datasets,

attention is shifting from algorithm development toward ensuring that AI systems can be deployed safely and reliably in routine clinical practice [10, 17].

External validation on independent datasets is widely considered as an essential step toward establishing model generalizability and has become standard practice in medical imaging AI [19, 23]. However, successful external validation does not necessarily imply deployment readiness. Differences between development and deployment environments—including patient populations, disease prevalence, imaging devices, acquisition protocols, and clinical workflows—may substantially affect model behaviour once deployed [9, 15]. Consequently, models validated across multiple international datasets may still fail when transferred to previously unseen healthcare settings. This challenge is particularly relevant in African healthcare systems, where medical imaging datasets remain underrepresented in the development and validation of AI models [1, 7, 16]. As a result, AI systems developed elsewhere might be considered for local deployment despite limited evidence of their performance in African clinical contexts. Cross-context evaluation on representative local data therefore provides an important opportunity to assess whether existing validation evidence is sufficient to support deployment decisions.

Diabetic retinopathy (DR) screening provides a representative use case. Recent deep learning systems have demonstrated strong performance across have been evaluated across multiple retinal imaging datasets [3, 22], yet African populations remain largely absent from commonly used evaluation benchmarks [16]. Whether such externally validated systems maintain their performance in African clinical settings remains largely unexplored. In this work, we investigate this question by evaluating two publicly released DR screening models—an interpretable sparse BagNet and a conventional ResNet-50—originally trained on the Kaggle DR dataset and externally validated across ten public datasets [3]. Without retraining or adapting the released models, we assess their performance on a private retinal fundus dataset collected in Kenya that was not included in the original evaluation. Beyond conventional performance metrics, we analyze operating characteristics across decision thresholds and examine representative prediction successes and failures using model interpretability.

Rather than proposing a new AI model, this study examines whether extensive external validation is sufficient to establish deployment readiness in a previously unseen African clinical setting. Using the Kenyan cohort as a practical case study, we discuss how local cross-context evaluation can complement conventional external validation and support more responsible deployment of medical imaging AI. Our contributions are threefold:

- We independently evaluate two externally validated DR screening models on an African fundus dataset not included in the original validation study.
- We characterize deployment-relevant performance through quantitative evaluation, threshold analysis, and qualitative interpretability assessment.
- We discuss how local cross-context evaluation can inform deployment readiness and responsible adoption of medical imaging AI in underrepresented healthcare settings.

2 Related Work

External validation of medical imaging AI. External validation is widely used to assess whether medical AI models generalize beyond their development datasets. In diabetic retinopathy screening, several studies have reported evaluations across multiple retinal imaging datasets with promising cross-dataset performance [3, 22, 26]. However, successful validation on existing benchmarks does not necessarily reflect performance in future deployment settings with different populations or clinical conditions [1, 9]

From model validation to deployment readiness. Recent work differentiates model validation from real-world deployment [15]. Beyond predictive performance, responsible deployment requires consideration of workflow integration, monitoring, human oversight, and evaluation within the intended clinical environment [17, 19]. Practical approaches for assessing deployment readiness before implementation, however, remain limited [6].

Medical imaging AI in African healthcare contexts. Medical AI could improve access to diagnostic services across Africa, yet African populations remain underrepresented in many AI development and validation datasets [1, 7]. Consequently, the transferability of existing models to African clinical settings remains uncertain [24], highlighting the importance of local cross-context evaluation before deployment.

3 Materials and Methods

3.1 Original models

We evaluated two publicly available DR screening models released by *Djoumessi et al.* [3] for binary classification (grade $\{0\}$ vs. $\{1, 2, 3, 4\}$). The original study compared an inherently interpretable sparse BagNet [2,4] trained with a sparsity constraint [4] against a conventional ResNet-50 [14]. Both models were trained on the Kaggle DR dataset [5], comprising approximately 73% healthy and 27% referable DR images, and externally validated on ten public datasets spanning Europe, Asia, and the Americas, demonstrating promising cross-dataset generalization [3]. The released model weights were used without modification.

3.2 Kenya dataset

The Kenya dataset is a private collection of retinal fundus images from the Nakuru Eye Disease Study [13], graded for diabetic retinopathy according to the International Clinical Diabetic Retinopathy (ICDR) severity scale [25]. A subset comprising $2,185$ participants and $8,703$ gradable fundus images was available for this study. The released dataset provides separate left- and right-eye DR grades, but the correspondence between individual fundus images and their associated eye labels is unavailable.

Table 1. Summary of dataset characteristics and model performance. "Origin" indicates the country of data collection, and "Prev." the prevalence of diabetic retinopathy.

Dataset	Origin	Number of Images			Prev	ResNet-50			sparse BagNet		
		All	Healthy	DR		AUC	Sens	Spec	AUC	Sens	Spec
Kaggle	USA	6,956	5,118	1,838	26%	0.935	0.765	0.967	0.904	0.706	0.978
IDRiD	India	512	168	348	68%	0.864	0.963	0.714	0.878	0.951	0.768
E-Ophtha	France	434	260	174	40%	0.972	0.960	0.912	0.944	0.920	0.892
FGA-DR	UAE	1,841	101	1,740	94%	0.816	0.764	0.743	0.789	0.807	0.594
DDR	China	12,513	6,265	6,248	50%	0.963	0.801	0.973	0.926	0.668	0.980
DR2	Brazil	445	300	145	32%	0.866	0.669	0.977	0.922	0.662	0.983
APTOS	USA	3,662	1,805	1,857	51%	0.972	0.942	0.956	0.995	0.982	0.965
FCM-UNA	Paraguay	757	187	570	75%	0.967	0.840	0.989	0.935	0.698	0.989
Messidor-1	France	1,200	546	654	54%	0.954	0.853	0.941	0.943	0.833	0.960
Messidor-2	France	1,744	1,017	727	42%	0.925	0.794	0.893	0.876	0.751	0.887
DIARETDB1	Finland	89	05	84	94%	0.819	0.667	1.000	0.943	0.798	1.000
Nakuru	Kenya	8,162	7,774	388	05%	0.715	0.357	0.977	0.710	0.257	0.979

To avoid ambiguous image-label assignments, the evaluation was restricted to participants with identical left- and right-eye grades, resulting in a cohort of 2,050 participants and 8,162 retinal images, with approximately 95% healthy and 5% DR images. To ensure a fair comparison with the original study, all images were preprocessed using the same pipeline as in [3]: circular field-of-view cropping [20], resizing to 512×512 pixels, and normalization using the Kaggle training-set statistics.

3.3 Evaluation protocol

The released sparse BagNet and ResNet-50 models were evaluated on the Kenya dataset using the same binary DR classification task and preprocessing pipeline as the original study [3]. No retraining, fine-tuning, threshold optimization, or domain adaptation was performed. Performance was assessed using the area under the receiver operating characteristic curve (AUC), sensitivity, and specificity—reflecting clinically relevant screening performance. To further characterize a deployment-relevant setting, we additionally evaluated probability calibration using the Expected Calibration Error (ECE), analyzed operating characteristics across decision thresholds, and qualitatively examined representative predictions using the sparse BagNet evidence maps.

4 Results

4.1 Reproducibility

We first reproduced the evaluation reported by *Djoumessi et al.* [3] using the publicly released models and the original evaluation protocol. The reproduced performance (Tab. 1) closely matches the published results across all previously reported external validation datasets, confirming faithful implementation.

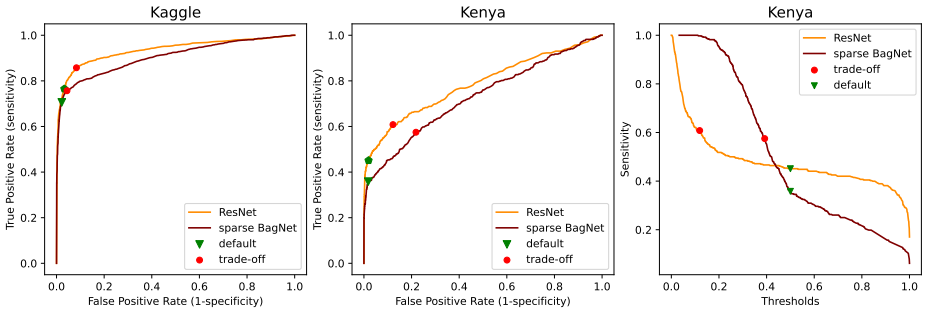

Fig. 1. Operating characteristics before deployment. ROC curves compare the discriminative performance of the ResNet-50 and sparse BagNet on the Kaggle and Kenyan datasets. The rightmost column illustrate the sensitivity across decision thresholds for the Kenyan cohort. Red markers correspond to the default threshold and green markers to the threshold maximizing Youden's J statistic.

4.2 Cross-context evaluation in the Kenyan clinical cohort

We next evaluated the models on the Nakuru dataset (Tab. 1. last row) using the same evaluation protocol as in the original study. While both models achieved strong performance across the previously reported internal and external datasets, performance on the Kenyan cohort was consistently lower.

Both models retained moderate discriminative performance on the Kenyan dataset, with an AUC of 0.715 (95% CI: [0.692 − 0.739]) for ResNet-50 and 0.710 (95% CI: [0.687 − 0.733]) for the sparse BagNet. These values were below those reported on the ten external validation datasets, where AUC ranged from 0.816 to 0.972 for ResNet-50 and from 0.789 to 0.995 for the sparse BagNet. The largest change was observed in sensitivity. On the Kenyan dataset, sensitivity decreased to 0.357 for ResNet-50 and 0.257 for the sparse BagNet, compared with values ranging from approximately 0.66 to 0.98 across the previously evaluated datasets. In contrast, specificity remained high for both models (0.977 [0.974 − 0.980] and 0.979 [0, 976 − 0.982], respectively), consistent with the values observed on several external datasets. Despite their different architectures, the two models exhibited similar performance trends on the Kenyan cohort, including reduced AUC and substantially lower sensitivity while maintaining high specificity. Extended results with confidence intervals are provided in Appendix A.

4.3 Operating Characteristics for Local Deployment

To further investigate the reduced sensitivity observed on the Kenyan cohort, we first assessed model calibration. The ResNet-50 remained well calibrated in both datasets, with the Expected Calibration Error (ECE) increasing only slightly from 0.021 on Kaggle to 0.040 on the Kenyan dataset. The sparse BagNet

exhibited higher ECE values on both datasets (0.224 and 0.280, respectively), although only a modest increase was observed on the Kenyan cohort.

We next investigated whether the operating threshold used in the original study remained appropriate for the Kenyan cohort. Figure 1 shows the ROC curves together with the default operating points and the threshold maximizing Youden's J statistic [21]. For the ResNet-50, the default threshold achieved a specificity of 0.981 and a sensitivity of 0.451. Selecting the threshold maximizing Youden's J statistic ($\tau = 0.119$) increased sensitivity to 0.608 while specificity decreased to 0.878. For the sparse BagNet, sensitivity increased from 0.358 to 0.575 after threshold adjustment ($\tau = 0.392$), followed by a substantial reduction in specificity from 0.982 to 0.781. This larger trade-off suggests that the interpretable model required a greater compromise between sensitivity and specificity in the Kenyan cohort. Overall, threshold optimization increased sensitivity for both models, although neither model reached the sensitivity observed on several previously evaluated external validation datasets.

4.4 Qualitative Interpretation for Deployment Assessment

Finally, we qualitatively examined the evidence maps produced by the inherently interpretable sparse BagNet for representative true-positive (TP), false-positive (FP), and false-negative (FN) predictions from the Kenyan dataset (Fig. 2). The sparse BagNet is a patch-based architecture that aggregates local evidence into the final prediction, allowing individual image regions to contribute directly to the classification decision [2, 4]. Evidence maps were extracted from the final evidence layer before aggregation. Bounding boxes (33×33 pixels), corresponding to the model receptive field, were drawn around the highest-relevant regions, and the top-12 image patches with the strongest evidence were extracted.

For correctly classified DR images (first row), the evidence maps highlighted several localized regions with strong positive evidence. Visual inspection of the corresponding patches revealed visible retinal abnormalities consistent with diabetic retinopathy, including microaneurysms, haemorrhages, and exudate-like structures. However, for representative false-positive predictions (second row), the evidence maps also identified multiple localized regions with positive disease evidence despite the image being labelled as non-diabetic retinopathy. Several of the extracted patches exhibited retinal structures visually similar to diabetic retinopathy lesions. For representative false-negative cases (third row), the evidence map exhibited predominantly negative evidence across the retinal image, with only limited localized positive evidence. In most cases, the evidence maps do not contain positive disease-activation regions, indicating the absence of retinal abnormalities. Additional examples of true-positive, true-negative, false-positive, and false-negative predictions are provided in Appendix B.

5 Discussion

Overall, our results suggest that strong external validation does not necessarily imply deployment readiness in a previously unseen clinical setting. Although

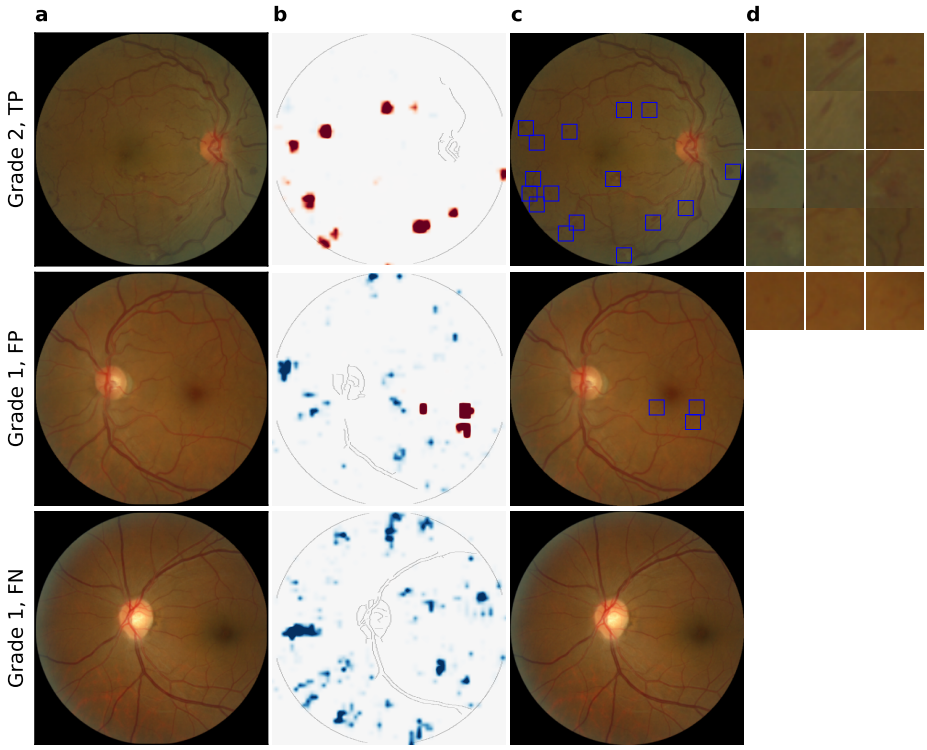

Fig. 2. Representative sparse BagNet explanations on the Kenyan dataset. From left to right: input image, evidence map, localized evidence regions, and the twelve highest-scoring evidence patches for representative true-positive, false-positive and false-negative predictions.

both DR screening models generalized well across ten publicly available datasets, evaluation on the Kenyan cohort revealed a consistent reduction in sensitivity, entirely driven by Grade 1 diabetic retinopathy, with 78% of Grade 1 images misclassified as healthy at the default threshold. No false negatives were observed for Grades 2–4, indicating that the observed degradation primarily concerned subtle early lesions rather than more advanced disease. Together with the threshold, calibration, and qualitative analyses, these findings illustrate how local cross-context evaluation can provide deployment-relevant evidence beyond conventional external validation.

From external validation to deployment assessment. The similar behaviour of the ResNet-50 and sparse BagNet suggests that the observed reduction was not specific to a single architecture, but reflects limitations in transferring benchmark validation results to a new clinical context. Our results support a practical pre-deployment assessment pathway in which external validation is followed by local evaluation of frozen models, calibration assessment, operating-

threshold analysis, grade-specific error characterization, and qualitative model auditing before a deployment decision is made. The threshold analysis further showed that local adjustment can improve sensitivity, although at the cost of specificity and without recovering the performance observed on previous external datasets. Thus, threshold selection can address part of the deployment gap but does not eliminate the need for local evaluation. This complements recommendations that trustworthy medical AI should be assessed beyond aggregate predictive performance and within its intended context of use [10, 17].

Local model auditing and deployment decisions. The qualitative analysis illustrates how interpretable models can support local auditing by making prediction-relevant image regions available for visual inspection. In this study, these explanations were used for model auditing rather than clinical validation: their interpretation was visual and was not independently assessed by ophthalmologists. Nevertheless, the combination of grade-specific errors and localized evidence provides a practical way to identify cases requiring further investigation before deployment. More broadly, local auditing can complement quantitative evaluation by helping identify whether errors arise predominantly from subtle disease patterns, image characteristics, or potentially ambiguous annotations. Such context-specific evidence is particularly relevant for challenging screening tasks in underrepresented healthcare settings, where assumptions based solely on previously validated datasets may not adequately reflect local populations or clinical conditions [1, 7, 15].

Limitations and future directions. This study evaluated a single Kenyan cohort and two publicly released models, and therefore represents a cross-context case study rather than a comprehensive deployment-readiness framework. The models were intentionally evaluated without retraining or adaptation to emulate pre-deployment assessment of an existing AI system. The analysis also focused on binary early DR screening, while incomplete metadata limited the possibility of detailed demographic or clinical subgroup analyses. Furthermore, the qualitative explanations were not independently reviewed by ophthalmologists and should therefore not be interpreted as evidence of clinical correctness. Although confidence intervals were estimated for the reported performance measures, formal hypothesis testing was not performed. Future work should validate these findings across additional African cohorts and prospective clinical workflows, incorporate clinician-reviewed error analysis and appropriate statistical comparisons [3], and investigate adaptation strategies such as fine-tuning with small target-site samples [11]. Such studies could determine whether local adaptation can mitigate observed performance degradation while maintaining independent evaluation of deployment performance.

6 Conclusion

We presented an independent cross-context evaluation of two externally validated DR screening models on a Kenyan retinal imaging cohort. Despite strong

performance across previous external datasets, both models showed reduced sensitivity in the new clinical context (Kenya dataset), driven entirely by missed Grade 1 cases. Threshold adjustment partially improved sensitivity, while calibration, grade-specific error analysis, and qualitative auditing provided complementary evidence for assessing model behaviour in the new context.

Rather than proposing a new model or a complete deployment-readiness framework, this study highlights local cross-context evaluation as an important step between external validation and clinical deployment. We argue that deployment decisions should be supported by context-specific evidence, including local performance assessment, operating-threshold evaluation, and qualitative model auditing before adoption in routine clinical practice.

Acknowledgments. This project was supported by the Hertie Foundation and by the Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy with the Excellence Cluster 2064 "Machine Learning – New Perspectives for Science", and a regular grant (project number 571331899).

Disclosure of Interests. The authors declare no competing interests.

References

1. Alaran, M.A., Lawal, S.K., Jiya, M.H., Egya, S.A., Ahmed, M.M., Abdulsalam, A., Haruna, U.A., Musa, M.K., Lucero-Prisno III, D.E.: Challenges and opportunities of artificial intelligence in african health space. Digital health **11**, 20552076241305915 (2025)
2. Brendel, W., Bethge, M.: Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. International Conference on Learning Representations (2019)
3. Djoumessi, K., Huang, Z., Kühlewein, L., Rickmann, A., Simon, N., Koch, L.M., Berens, P.: An inherently interpretable ai model improves screening speed and accuracy for early diabetic retinopathy. PLOS Digital Health **4**(5), e0000831 (2025)
4. Djoumessi, K., Ilanchezian, I., Kühlewein, L., Faber, H., Baumgartner, C.F., Bah, B., Berens, P., Koch, L.M.: Sparse activations for interpretable disease grading. In: Medical Imaging with Deep Learning (2023)
5. Dugas, E., Jared, J., Cukierski, W.: Diabetic retinopathy detection (2015), `https://kaggle.com/competitions/diabetic-retinopathy-detection`
6. El Miayar, O., Berrado, A.: Why health recommender systems struggle to reach clinical practice: A lifecycle-oriented systematic review. iScience **29**(6) (2026)
7. Ephraim, R.K.D., Kotam, G.P., Duah, E., Ghartey, F.N., Mathebula, E.M., Mashamba-Thompson, T.P.: Application of medical artificial intelligence technology in sub-saharan africa: prospects for medical laboratories. Smart Health **33**, 100505 (2024)
8. Ferrer, L., Riera, P.: Confidence intervals for evaluation in machine learning, `https://github.com/luferrer/ConfidenceIntervals`
9. Finlayson, S.G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I.S., Saria, S.: The clinician and dataset shift in artificial intelligence. New England Journal of Medicine **385**(3), 283–286 (2021)

10. Gonzalez-Gonzalo, C., Thee, E.F., Klaver, C.C., Lee, A.Y., Schlingemann, R.O., Tufail, A., Verbraak, F., Sánchez, C.I.: Trustworthy ai: Closing the gap between development and integration of ai systems in ophthalmic practice. Progress in retinal and eye research **90**, 101034 (2022)
11. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE Transactions on Biomedical Engineering **69**(3), 1173–1185 (2021)
12. 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 **316**(22), 2402–2410 (2016)
13. Hansen, M.B., Abràmoff, M.D., Folk, J.C., Mathenge, W., Bastawrous, A., Peto, T.: Results of automated retinal image analysis for detection of diabetic retinopathy from the nakuru study, kenya. PloS one **10**(10), e0139148 (2015)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
15. Koch, L.M., Baumgartner, C.F., Berens, P.: Distribution shift detection for the postmarket surveillance of medical ai algorithms: a retrospective simulation study. NPJ Digital Medicine **7**(1), 120 (2024)
16. Krzywicki, T., Brona, P., Zbrzezny, A.M., Grzybowski, A.E.: A global review of publicly available datasets containing fundus images: characteristics, barriers to access, usability, and generalizability. Journal of Clinical Medicine **12**(10), 3587 (2023)
17. Lekadir, K., Feragen, A., Fofanah, A., Frangi, A., Buyx, A., Emelie, A., Lara, A., Porras, A., Chan, A., Navarro, A., et al.: Future-ai: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. arxiv. arXiv preprint arXiv:2309.12325 (2023)
18. Liu, X., Faes, L., Kale, A.U., Wagner, S.K., Fu, D.J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., et al.: A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The lancet digital health **1**(6), e271–e297 (2019)
19. Marconi, L., Cabitza, F.: Show and tell: a critical review on robustness and uncertainty for a more responsible medical ai. International journal of medical informatics **202**, 105970 (2025)
20. Mueller, S., Heidrich, H., Koch, L.M., Berens, P.: fundus circle cropping. `https://doi.org/10.5281/zenodo.10137935`, `https://github.com/berenslab/fundus_circle_cropping`
21. Ruopp, M.D., Perkins, N.J., Whitcomb, B.W., Schisterman, E.F.: Youden index and optimal cut-point estimated from observations affected by a lower limit of detection. Biometrical Journal: Journal of Mathematical Methods in Biosciences **50**(3), 419–430 (2008)
22. Selvachandran, G., Quek, S.G., Paramesran, R., Ding, W., Son, L.H.: Developments in the detection of diabetic retinopathy: a state-of-the-art review of computer-aided diagnosis and machine learning methods. Artificial intelligence review **56**(2), 915–964 (2023)
23. Suleman, M.U., Mursaleen, M., Khalil, U., Saboor, A., Bilal, M., Khan, S.A., Subhani, M.A., Hussnain, M.A., Tabassum, S.N., Tahir, M.: Assessing the generalizability of artificial intelligence in radiology: a systematic review of performance across different clinical settings. Annals of Medicine and Surgery **87**(12), 8803–8811 (2025)

24. Tadesse, A.Z., Zemede, B.L., Shimber, E.T., Kebede, S., Fekele, Y., Nsengiyumva, E., Sultan, M., Mohamud, M.F.Y., Mdundo, W., Manirafasha, L.A., et al.: Perception and challenges of artificial intelligence (ai) in emergency medicine: A multi-country study in sub-saharan africa. African Journal of Emergency Medicine **16**(3), 100978 (2026)
25. Wong, T.Y., Sun, J., Kawasaki, R., Ruamviboonsuk, P., Gupta, N., Lansingh, V.C., Maia, M., Mathenge, W., Moreker, S., Muqit, M.M., et al.: Guidelines on diabetic eye care: the international council of ophthalmology recommendations for screening, follow-up, referral, and treatment based on resource settings. Ophthalmology **125**(10), 1608–1622 (2018)
26. 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 **622**(7981), 156–163 (2023)