

Diagnostic efficacy of a novel Diabetic Retinopathy AI algorithm in real-world undilated retinal images

identifeye HEALTH, University of Oulu, Optomed Plc

September 28, 2026

Abstract

Accessible and accurate screening for diabetic retinopathy (DR) can help prevent avoidable vision loss, particularly where access to standard retinal imaging is limited. Automated portable fundus imaging platforms, along with artificial intelligence (AI) can bring screening closer to patients, and support faster, more scalable workflows. We evaluated the identifeye HEALTH DR AI algorithm on an independent dataset of 156 patients from a dataset collected by Oulu University Hospital. The algorithm achieved 93.33% sensitivity, 92.31% specificity, and 96.18% gradability in the same dataset outperforms existing models as described in Kubin et al. 2024.⁴

1 Introduction

Diabetic retinopathy (DR) is impacting an estimated 9.6 million patients in the US (as of 2021)¹ and 103 million patients worldwide^{2,3}, with prevalence expected to increase substantially over the next decade. Regular DR screening enables earlier detection when behavioral modifications and medical interventions can significantly change the outcomes, yet access in many settings remains limited.

Many AI-based DR screening and classification algorithms have been developed, but their performance in real-world conditions varies substantially, when compared on shared datasets.⁴⁻⁶ Similar differences in diagnostic performance have been noted in other studies⁷, which identified differences in training populations, image quality, and representation of comorbidities as possible contributors. This underscores the need to evaluate DR algorithms using imaging and patient populations that truly reflect their intended clinical use.

Here, we evaluate the identifeye HEALTH DR algorithm on a real-world dataset acquired with a handheld retinal camera and compare its performance with published results for 21 algorithms evaluated on the same dataset. We assess performance across three clinically relevant classification tasks - more-than-mild DR (mtmDR), vision threatening DR (vtDR), and any DR; and examine how macula-centered, optic disc-centered, and combined dual-field inputs affect performance. This analysis evaluates both the diagnostic capability of the algorithm and the contribution of retinal field coverage to robust AI-enabled screening.

2 Methods

2.1 System Architecture

The identify DR classification system consists of three AI-based modules: (1) an image quality assessment (IQA) model to evaluate images appropriate for interpretation, (2) a localizer model to identify the retinal field (macula or optic disc) and eye laterality (right or left), and (3) A DR classification model.

The DR classifier assigns one of the five International Clinical Diabetic Retinopathy Disease (ICDR) severity scale categories: no DR, mild nonproliferative DR (NPDR), moderate NPDR, severe NPDR, and proliferative DR (PDR).⁸ The model also predicts the presence of individual DR-associated lesions.

2.2 Model Development

Models were developed using images from the EyePACS⁹ dataset. Images were curated by imaging field, image quality, and imaging device. Data was split on patient ID to prevent data leakage between subsets (Figure 1). To evaluate generalization across imaging devices we ensure some non-overlapping devices across the train, validation, and test subsets.

EyePACS human reader annotations include per-eye DR severity and the presence or absence of 11 lesion types: microaneurysms, hemorrhages, cotton wool spots, hard exudates, intraretinal microvascular abnormalities (IRMA), neovascularization or fibrous proliferation (NVFP), diabetic macular edema (DME), laser scarring, venous beading, and preretinal and vitreous hemorrhage (PRHVH). As a data quality check, DR severity derived from the lesions is verified on a per-eye basis by EyePACS's grading template, resulting in a 96% agreement.

The identify DR classifier uses a modified EfficientNet-V2¹⁰ convolutional neural network trained to predict the five-level DR severity and lesion presence per image. Experimentally, we found that multi-task training of multi-class DR classification and multi-label lesion identification improved the overall DR classification accuracy.

2.3 DR endpoints and prediction aggregation

Three classification endpoints were considered:

- Less-than-mild DR: ICDR 0 - 1
- More-than-mild DR (mtmDR): ICDR grade 2 or higher
- Vision-threatening DR (vtDR): ICDR grade 3 or higher

Cases predicted as ICDR grade 2 or higher are determined to be referable for a retinal specialist evaluation.

For single-field analyses, predictions were generated separately using the macula-centered and optic disc-centered images. For dual-field analysis, the maximum prediction of macula- and optic disc-centered image-level predictions was used. Both eye-level and subject-level mtmDR probabilities were determined heuristically. Eye-level scoring was determined by the maximum DR probability across available images per eye. Subject-level scoring followed the ruleset:

- A subject was classified as referable if at least one eye was referable.
- A subject was classified as non-referable if both eyes were non-referable.

- A subject was classified as ungradable if neither eye was referable and at least one eye was ungradable.

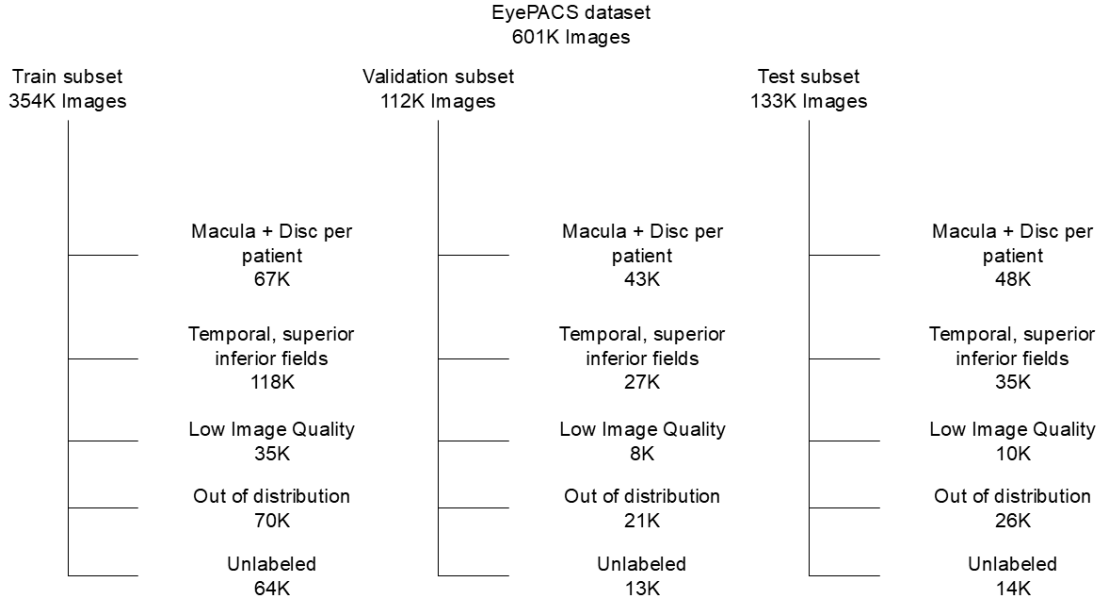


Figure 1: EyePACS data breakdown

2.4 External Evaluation Dataset

External performance was evaluated using a dataset collected by Oulu University Hospital where retinal fundus images were acquired using the Optomed Aurora handheld retinal camera⁴. The dataset included 156 patients and a total of 624 retinal images acquired with Optomed Aurora. A small subset of images was excluded because they were recorded as grayscale. Patients in this dataset contained comorbidities such as age-related macular degeneration (AMD), branch retinal vein occlusion (BRVO), and central retinal vein occlusion (CRVO).

The Oulu dataset was not used for AI model development, hyperparameter selection, threshold selection, or calibration. The identifeye DR model and all decision rules were fixed before evaluation on the dataset. For five-class DR predictions, the ICDR level with the highest probability was selected (using the “argmax” operation). Predicted grades of moderate NPDR or higher were categorized as referable, whereas no DR and mild NPDR were categorized as non-referable. Performance was calculated at both per-eye and per-subject basis. We also analyzed single-field versus dual-field performance.

3 Results

Table 1 shows the identifeye DR model’s performance on the binary DR task (referable/non-referable DR) as well as the gradeability (proportion of images not rejected by the IQA module for insufficient quality). A total of 8 eyes out of 314 eyes were rejected and removed from analysis since they were grayscale and/or retinal landmarks (macula and optic disc) were missing.

Table 1: Performance of the identifeye DR AI model on binary DR classification

Binary DR	Sensitivity	Specificity	PPV	NPV	Gradeability
Per-subject	0.9333	0.9231	0.9545	0.8889	0.9618
Per-eye	0.9074	0.9495	0.9495	0.9074	0.9745

Table 2 shows the identifeye DR model's performance on classifying vision-threatening DR (vtDR). vtDR is defined as DR grades 3 or 4 that require immediate clinical attention. The model provides referrals for 97.6% of eyes with vtDR.

Table 2: Performance of the identifeye DR AI model on vision-threatening DR (vtDR) classification

vtDR	Sensitivity	Specificity	PPV	NPV
Per-subject	0.8800	0.8571	0.5500	0.9730
Per-eye	0.9048	0.8939	0.5758	0.9833

Table 3 shows the identifeye DR model's performance on any level of DR, defined as DR grade of 1 or above. Being able to predict DR 1 cases - defined as those with the presence of only microaneurysms - can be beneficial for early detection.

Table 3: Performance of the identifeye DR AI model on any DR classification

Any-DR	Sensitivity	Specificity	PPV	NPV
Per-subject	0.9286	0.9012	0.8904	0.9360
Per-eye	0.9225	0.9435	0.9225	0.9435

Table 4 evaluates the dependency of DR model's classification performance on retinal field. This evaluation included sets of eyes for which both the macula and optic disc fields were imaged. Macula-centered performance suffers slightly compared with disc-centered and macula-and-disc performances.

Tables 5 and 6 show the DR model's vtDR and any-DR performance on macula-centered, disc-centered, and dual field imaging scenarios.

Table 4: Performance of the identifeye DR AI model on binary DR based on included retinal fields

	Fields	Sensitivity	Specificity	Gradeability
Per-eye (common set)	Macula + Disc	0.9074	0.9495	0.9745
	Macula-centered	0.8980	0.9556	0.8854
	Disc-centered	0.9167	0.9553	0.8758
Per-subject	Macula + Disc	0.9333	0.9231	0.9618
	Macula-centered	0.9474	0.9221	0.8535
	Disc-centered	0.9474	0.9375	0.8726

Table 5: Performance of the identifeye DR AI model on vtDR classification based on included retinal fields

	Fields	Sensitivity	Specificity
Per-eye (common set)	Macula + Disc	0.9047	0.8939
	Macula-centered	0.8529	0.9180
	Disc-centered	0.8974	0.9067
Per-subject	Macula + Disc	0.8800	0.8571
	Macula-centered	0.9000	0.8772
	Disc-centered	0.8750	0.8673

Table 6: Performance of the identifeye DR AI model on any-DR classification based on included retinal fields

	Fields	Sensitivity	Specificity
Per-eye (common set)	Macula + Disc	0.9225	0.9435
	Macula-centered	0.9138	0.9630
	Disc-centered	0.8974	0.9620
Per-subject	Macula + Disc	0.9286	0.9012
	Macula-centered	0.9538	0.9275
	Disc-centered	0.9402	0.9286

On this dataset binary mtmDR performance on macula-only, disc-only, and dual-field subsets is similar. We hypothesize that this dataset is too small to resolve the performance differences in these subsets, but assume that considering more fields should lead to better DR screening performance as more lesions can be captured. Appendix A lists some discordant predictions when the retinal field is artificially restricted (i.e. excluding a field for evaluation even when it is present). While in most cases the discordant predictions differ only by a single DR grade, occasionally the difference crosses the referability threshold, indicating that dual-field imaging should improve DR classification performance compared with single-field.

3.1 Comparison with previously evaluated algorithms

For binary DR classification the identifeye DR model outperformed 21 AI models previously evaluated by Oulu on the same dataset^{4,11}, achieving the highest Youden index (0.856) while maintaining 98% gradeability (Table 7).

Table 7: AI Agreement Performance Metrics

AI Model Ranking	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Youden's index
Identifeye DR AI	93.3	92.3	95.4	88.8	0.856
1	92.3	85.0	96.9	94.4	0.82
2	91.0	85.0	94.8	91.1	0.80
3	85.9	91.7	82.3	79.7	0.74
4a	87.8	76.7	94.8	90.2	0.71
4b	85.9	83.1	87.6	80.3	0.71
6a	85.3	81.7	87.5	80.3	0.69
6b	86.5	76.7	92.7	86.8	0.69
8	84.0	83.3	84.4	78.1	0.68
9a	82.1	83.3	81.3	84.3	0.65
9b	84.0	76.7	88.5	80.7	0.65

A study by Aireen AI reported a Youden index of 0.862⁶ for any-DR classification task, distinguishing no DR from any DR level (grade 0 vs grades 1 - 4). In comparison, the results reported above are for binary mtmDR classification, distinguishing grades 0 - 1 from 2 - 4. When evaluated on the same any-DR task, the identifeye DR model achieved a sensitivity of 92.25%, a specificity of 94.35%, and a Youden index of 0.866.

Image quality is critical for reliable real-world retinal screening and image interpretation, whether the analysis is performed by clinicians, AI algorithms, or a combination. The principle extends beyond DR to virtually all screening applications. By increasing the image quality threshold for acceptable images, DR performance can be improved further, although this comes at the expense of imageability, as more images are classified as ungradable.

Figure 2 illustrates this tradeoff. As the image quality threshold is increased, imageability decreases, while sensitivity and specificity for binary classification improve. Youden's index is shown on the left and harmonic mean of sensitivity and specificity is shown on the right. The improvement is even more pronounced when performance is evaluated on a per-subject basis. Since the dataset size is small (N=314 eyes), fluctuations in the curves are expected.

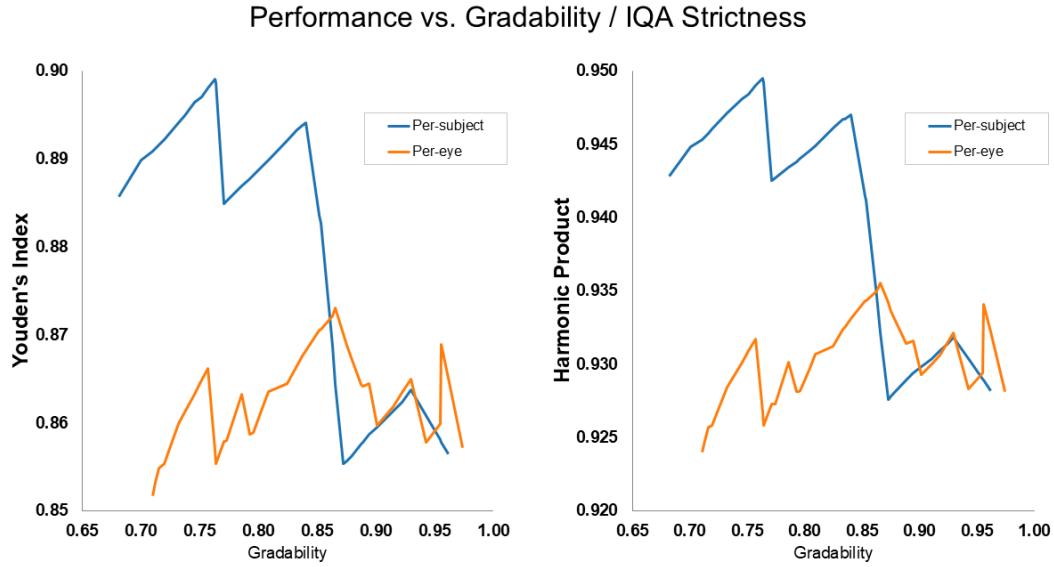


Figure 2: identefeye DR Performance versus Imageability. Left: Youden’s Index. Right: Harmonic Mean of Specificity and Sensitivity.

While reduced imageability may seem unfavorable from a workflow point of view, image capture automation can help mitigate this problem by consistently capturing higher-quality images with fewer image capture attempts. Applying more stringent image quality thresholds improves diagnostic accuracy and ultimately enhances screening performance.

4 Conclusion

In this study, we evaluated a new robust DR classification algorithm developed by identefeye HEALTH on an independent, real-world dataset acquired with a handheld retinal fundus camera at Oulu University Hospital. The identefeye DR AI-based classification framework demonstrates state-of-the-art performance, achieving higher performance than 21 existing algorithms previously evaluated on the same dataset, without retraining, calibration, or threshold adjustment on the evaluation data.

This analysis also highlights the importance of image acquisition workflow. While single-field retinal photography offers workflow efficiency, comparative analysis shows that macula-centered and optic-disc-centered images provide complementary diagnostic information. Restricting analysis to either field alone can exclude clinically relevant lesions and reduce diagnostic accuracy. The findings reported here demonstrate both the diagnostic capability of the identefeye DR algorithm under real-world imaging conditions and the importance of retinal field coverage.

References

- [1] Lundeen EA, Burke-Conte Z, Rein DB, Wittenborn JS, Saaddine J, Lee AY, Flaxman AD. Prevalence of diabetic retinopathy in the US in 2021. *JAMA Ophthalmology*, 141(8):747–755, 2023. doi: 10.1001/jamaophthalmol.2023.2289.
- [2] Tan TE, Wong TY. Diabetic retinopathy: Looking forward to 2030. *Frontiers in Endocrinology*, 13:1077669, Jan 2023. doi: 10.3389/fendo.2022.1077669.

- [3] Towards Healthcare. Diabetic retinopathy market size to lead USD 18.97 Bn by 2035. Technical report, Towards Healthcare, Jan 2026. URL <https://www.towardshealthcare.com/insights/diabetic-retinopathy-market-sizing>. Accessed: 2026-01-22.
- [4] Kubin AM, Huhtinen P, Ohtonen P, Keskitalo A, Wirkkala J, Hautala N. Comparison of 21 artificial intelligence algorithms in automated diabetic retinopathy screening using handheld fundus camera. *Annals of Medicine*, 56(1):2352018, 2024. doi: 10.1080/07853890.2024.2352018.
- [5] Rudnicka AR, Shakespeare R, Chambers R, Bolter L, Anderson J, Fajtl J, Welikala RA, Barman SA, Olvera-Barrios A, Webster L, Mann S, Lee A, Remagnino P, Egan C, Owen CG, Tufail A; ARIAS Research Group. Automated retinal image analysis systems to triage for grading of diabetic retinopathy: a large-scale, open-label, national screening programme in England. *The Lancet Digital Health*, 7(11):100914, Nov 2025. doi: 10.1016/j.landig.2025.100914.
- [6] Huhtinen P, Kubin AM, Dvořák K, Sliva M, Bayer J, Hautala N. Real-world evaluation of artificial intelligence-based diabetic retinopathy screening using the Optomed Aurora handheld fundus camera. *Diabetes Technology & Therapeutics*, 27(12):1023–1025, Dec 2025. doi: 10.1177/15209156251369886.
- [7] Lee AY, Yanagihara RT, Lee CS, Blazes M, Jung HC, Chee YE, Gencarella MD, Gee H, Maa AY, Cockerham GC, Lynch M, Boyko EJ. Multicenter, head-to-head, real-world validation study of seven automated artificial intelligence diabetic retinopathy screening systems. *Diabetes Care*, 44(5):1168–1175, May 2021. doi: 10.2337/dc20-1877.
- [8] Wilkinson CP, Ferris FL 3rd, Klein RE, Lee PP, Agardh CD, Davis M, Dills D, Kampik A, Pararajasegaram R, Verdaguer JT; Global Diabetic Retinopathy Project Group. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*, 110(9):1677–1682, 2003. doi: 10.1016/S0161-6420(03)00475-5.
- [9] Cuadros J, Bresnick G. EyePACS: an adaptable telemedicine system for diabetic retinopathy screening. *Journal of Diabetes Science and Technology*, 3(3):509–516, 2009. doi: 10.1177/193229680900300315.
- [10] Tan M, Le QV. EfficientNetV2: Smaller models and faster training. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pages 10096–10106, 2021.
- [11] Huhtinen P, Kubin AM, Piila L, Hautala N. Performance and accuracy of 10 ai-based algorithms in diabetic retinopathy screening in a real-world clinical setting. *Acta Ophthalmologica*, 103(7):e532–e533, 2025. doi: 10.1111/aos.17542.

Appendix A: Eye-level Discordant predictions

Table 8: Discordant eye-level predictions: macula-and-disc versus disc-only.

Study ID	Eye	Macula and Disc	Disc Only
0059	left	4	0
0067	right	1	0
0106	right	1	0
0221	right	2	3
0223	left	1	0
0227	right	3	2
0243	left	1	0

Table 9: Discordant eye-level predictions: macula-and-disc versus macula-only.

Study ID	Eye	Macula and Disc	Macula Only
0074	left	2	0
0080	right	1	0
0086	left	2	1
0090	right	1	0
0095	left	1	0
0097	left	1	0
0104	right	1	0
0202	right	4	2
0208	left	3	2
0210	right	3	2
0213	right	4	2
0216	left	3	2
0222	left	3	2
0224	left	3	2
0228	left	4	3
0234	right	4	2
0237	right	4	3
0248	left	4	2