Supplementary Information

Performance of AI-based diabetic retinopathy screening is highly dependent on evaluation setting: A five-year, multi-framework study

Gwenolé Quellec[1*], Mathieu Lamard[2,1], Sarah Matta[2,1], Laurent Borderie[3], Philippe Zhang[2,1], Paola Ortega Saborio[3], Pierre Deman[3,4], Alexandre Le Guilcher[3], Ali Erginay[5], Anna-Maria Kubin[6,7], Nina Hautala[6,7], Petri Huhtinen[8], Béatrice Cochener[9,2,1], Pascale Massin[5]

[1]LaTIM UMR 1101, Inserm, Brest, 29200, France.
[2]Université de Bretagne Occidentale, Brest, 29200, France.
[3]Evolucare Technologies, Villers-Bretonneux, 80800, France.
[4]ADCIS S.A., Saint-Contest, 14280, France.
[5]Département d'ophtalmologie, Hôpital Lariboisière, AP-HP, Paris, 75010, France.
[6]Department of Ophthalmology, Oulu University Hospital, Oulu, 90220, Finland.
[7]University of Oulu, Oulu, 90220, Finland.
[8]Optomed, Oulu, 90230, Finland.
[9]Service d'Ophtalmologie, CHU Brest, Brest, 29200, France.

*Corresponding author(s). E-mail(s): gwenole.quellec@inserm.fr;

S1. AI system details

The evaluated system is a CE-marked artificial intelligence (AI) system designed for the automated analysis of color fundus photographs in diabetic retinopathy (DR) screening. It performs multiple tasks, including image quality assessment, eye laterality identification, detection of referable DR, detection of diabetic macular edema (DME), and grading of DR severity.

The system relies on ensembles of convolutional neural networks. Input images are preprocessed through resizing, cropping, and intensity normalization to reduce variability in illumination and image scale, as described elsewhere [1]. Several convolutional network architectures were investigated during development, including VGG, ResNet, Inception, DenseNet, and NASNet. The final ensembles retained only ResNet and Inception architectures (V3 and V4), which provided the best trade-off between performance and robustness.

Predictions are computed at the image level. When several photographs are available for the same eye, outputs are aggregated at the eye level by selecting the image with the highest predicted severity score. In addition to classification outputs, the system produces visual explanations highlighting image regions contributing to the decision [1].

Only the original version of the system, corresponding to the CE-marked release, was considered in this study. The system was not retrained or modified between the different evaluation settings described in the main manuscript.

S2. Training data

The system was trained on a large-scale annotated dataset derived from the OPHDIAT telemedicine screening program [2]. OPHDIAT collects retinal images from multiple clinical centers (40 centers as of 2017) under routine screening conditions. Each center is equipped with one of the following 45° digital non-mydriatic cameras: Canon CR-DGI or CR2 (Tokyo, Japan), Topcon TRC-NW6 or TR-NW400 (Rotterdam, The Netherlands). The full dataset included 763,848 images from 102,257 diabetic patients acquired between 2004 and 2017. A subset of images from an earlier data export (2008–2009 period) was used in initial developments.

Image grading was performed by certified ophthalmologists following standardized protocols, including quality control procedures such as double grading, and was consistent with French national guidelines for DR screening [3]. Quality assurance procedures included double grading of a subset of examinations. Images that received double grading were assigned to an internal test set of approximately 20,000 images from 10,378 eyes (including 9,734 eyes with sufficient image quality). Images from the remaining patients were used for training.

Regulatory and ethical compliance information is provided in the main Methods section.

S3. Detailed performance on Messidor-2

This section provides detailed benchmark results on Messidor-2 and on other internal evaluation tasks used during development of the AI system.

Two operating points were considered for referable DR detection on Messidor-2: a high-sensitivity setting (sensitivity: 99.0%, specificity: 88.2%) and a high-specificity setting (sensitivity: 95.3%, specificity: 92.1%). The high-sensitivity setting was retained in all experiments of the main manuscript for cross-setting comparisons because it is more consistent with the requirements of a screening workflow, where missed cases should be minimized.

The high-specificity setting is reported here to document the flexibility of the original system and to illustrate the trade-off between sensitivity and specificity that accompanies changes in decision threshold.

S4. Supplementary interpretation notes

The main manuscript focuses on performance variability across evaluation settings rather than on exhaustive reporting of benchmark results. The purpose of the present supplementary information is therefore to provide additional technical and benchmark-level detail without shifting the emphasis of the main article away from its central message: namely, that the observed performance of AI systems for DR screening depends strongly on the evaluation framework.

Benchmark performance should thus be interpreted as a conditional estimate of system behavior under a specific dataset and protocol, rather than as a sufficient description of real-world performance.

References

[1] Quellec, G., Charrière, K., Boudi, Y., Cochener, B. & Lamard, M. Deep image mining for diabetic retinopathy screening. *Med Image Anal* **39**, 178–193 (2017).

[2] Massin, P. *et al.* OPHDIAT: a telemedical network screening system for diabetic retinopathy in the Ile-de-France. *Diabetes Metab* **34**, 227–234 (2008).

[3] Massin, P. *et al.* Recommandations de l'ALFEDIAM pour le dépistage et la surveillance de la rétinopathie diabétique. *J Fr Ophtalmol* **20**, 302–310 (1997).

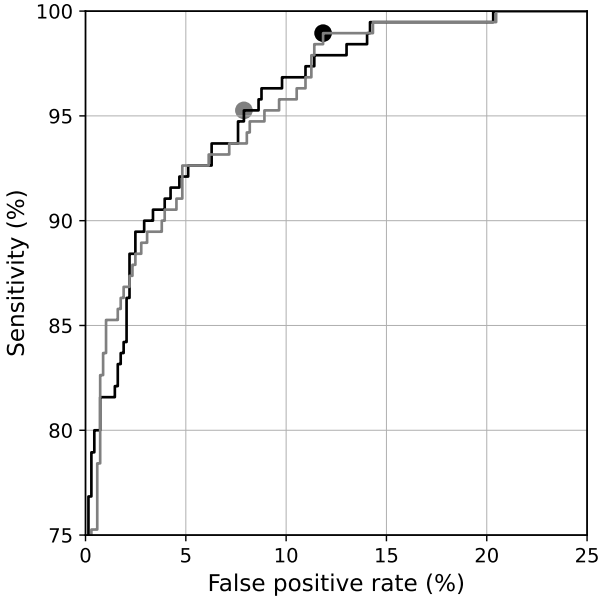

Fig. 1 Receiver operating characteristic (ROC) curves for referable diabetic retinopathy detection on Messidor-2. Two curves are shown, corresponding to models trained on different datasets: the full training set (2004–2017, black) and a subset (2008–2009, light gray). The operating point markers were intentionally color-inverted for visual clarity: the high-sensitivity operating point retained for subsequent evaluations is shown as a black marker on the light gray curve, whereas the high-specificity operating point is shown as a light gray marker on the black curve. The figure focuses on the clinically relevant region of low false positive rates and high sensitivity.

Table 1 Performance metrics of referable DR detection (on benchmark) and internal evaluation tasks (on internal test set). PDR: proliferative DR; NPDR: non-proliferative DR; DME: diabetic macular edema.

Task	**AUC**	**Sensitivity (%)**	**Specificity (%)**
Referable DR detection (2008-2009)	0.988	99.0	88.2
Referable DR detection (2004-2017)	0.989	95.3	92.1
Image quality assessment	0.971	92.4	90.0
Mild NPDR or more severe DR detection	0.970	90.3	93.0
Moderate NPDR or more severe DR detection	0.986	93.9	95.1
Severe NPDR or more severe DR detection	0.997	95.6	98.6
PDR detection	0.997	100.0	98.9
DME detection	0.994	96.8	96.3
Laterality identification	—	99.9% accuracy	—