PLOS DIGITAL HEALTH

RESEARCH ARTICLE

An inherently interpretable AI model improves screening speed and accuracy for early diabetic retinopathy

Kerol Djoumessi[1,2], **Ziwei Huang**[1,2], **Laura Kühlewein**[3], **Annekatrin Rickmann**[3,4], **Natalia Simon**[5], **Lisa M. Koch**[1,2,6], **Philipp Berens**[1,2*‡]

1 Hertie Institute for AI in Brain Health, University of Tübingen, Tübingen, Germany, **2** Tübingen AI Center, University of Tübingen, Tübingen, Germany, **3** University Eye Hospital, University of Tübingen, Tübingen, Germany, **4** Eye Clinic Sulzbach, Knappschaft Hospital Saar, Sulzbach, Germany, **5** Black Forest Eye Clinic, Endingen, Germany, **6** Department of Diabetes, Endocrinology, Nutritional Medicine and Metabolism UDEM, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland

‡ Current address: Hertie Institute for AI in Brain Health, University of Tübingen, Tübingen, Germany
* philipp.berens@uni-tuebingen.de

Check for updates

OPEN ACCESS

Citation: Djoumessi K, Huang Z, Kühlewein L, Rickmann A, Koch LM, Berens P (2025) An inherently interpretable AI model improves screening speed and accuracy for early diabetic retinopathy. PLOS Digit Health 4(5): e0000831. https://doi.org/10.1371/journal.pdig.0000831

Editor: Po-Chih Kuo, National Tsing-Hua University: National Tsing Hua University, TAIWAN

Received: January 07, 2025

Accepted: March 19, 2025

Published: May 12, 2025

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pdig.0000831

Copyright: © 2025 Djoumessi et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: The implementation of our sparse BagNet model is

Abstract

Diabetic retinopathy (DR) is a frequent complication of diabetes, affecting millions worldwide. Screening for this disease based on fundus images has been one of the first successful use cases for modern artificial intelligence in medicine. However, current state-of-the-art systems typically use black-box models to make referral decisions, requiring post-hoc methods for AI-human interaction and clinical decision support. We developed and evaluated an inherently interpretable deep learning model, which explicitly models the local evidence of DR as part of its network architecture, for clinical decision support in early DR screening. We trained the network on 34,350 high-quality fundus images from a publicly available dataset and validated its performance on a large range of ten external datasets. The inherently interpretable model was compared to post-hoc explainability techniques applied to a standard DNN architecture. For comparison, we obtained detailed lesion annotations from ophthalmologists on 65 images to study if the class evidence maps highlight clinically relevant information. We tested the clinical usefulness of our model in a retrospective reader study, where we compared screening for DR without AI support to screening with AI support with and without AI explanations. The inherently interpretable deep learning model obtained an accuracy of .906 [.900–.913] (95%-confidence interval) and an AUC of .904 [.894–.913] on the internal test set and similar performance on external datasets, comparable to the standard DNN. High evidence regions directly extracted from the model contained clinically relevant lesions such as microaneurysms or hemorrhages with a high precision of .960 [.941–.976], surpassing post-hoc techniques applied to a standard DNN. Decision support by the model highlighting high-evidence regions in the image improved screening accuracy for difficult decisions and improved screening speed. This shows that inherently interpretable deep learning models can provide clinical decision support while obtaining state-of-the-art performance improving human-AI collaboration.

available at GitHub (https://github.com/kdjoumessi/Sparse-BagNet_clinical-validation). The annotations performed for this study on selected Kaggle database images, the study data, and the analysis are available in the same GitHub repository.

Funding: This work was supported by a grant of the Hertie Foundation to PB, grants from the German Research Foundation (BE5601/8-1 to PB; Excellence Cluster 2064 "Machine Learning — New Perspectives for Science", project number 390727645 to PB), a grant from the Carl Zeiss Foundation ("Certification and Foundations of Safe Machine Learning Systems in Healthcare" to LK). Furthermore, the International Max Planck Research School for Intelligent Systems (IMPRS-IS) supported KD. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

AI systems designed to support clinical decision making use black-box deep learning models for many medical applications. This includes AI-based screening systems for diabetic retinopathy, a sight threatening complication of diabetes. This hinders clinical uptake of such methods, as clinicians and patients do not have a way to validate the AI systems decisions. Sometimes, post-hoc methods are used to generate heatmaps that supposedly explain the AI systems decision. However, these methods are problematic as the generated explanations do not reflect the actual decision-making process of the model and are prone to spurious correlations. In our paper, we take a big step forward for enabling trustworthy AI systems for supporting clinical decision making in screening for diabetic retinopathy: We introduce an inherently interpretable deep learning model which provides human-understandable explanations for its decisions. The model combines the power of deep learning with the interpretability of simpler models such as logistic regression by computing an explicit evidence map. This map forms the basis of the model's decisions, alleviating the issues of post-hoc techniques. We validate the clinical potential of this model for improving diabetic retinopathy screening showing highlighting regions with high disease evidence during clinical grading decreased the grading time significantly and improved grading accuracy for difficult borderline cases.

Introduction

Diabetic retinopathy (DR) screening has been one of the first successful use cases for artificial intelligence (AI) in medicine [1], promising fast, cost-effective access even where insufficient clinical personnel is available. By now, multiple AI systems have received regulatory clearance [2,3] and have been found useful to triage patients not requiring specialist attention and those with vision-threatening DR, potentially contributing to increased screening adherence [4].

However, current state-of-the-art models use black-box deep learning approaches to make referral decisions, providing clinicians only with limited binary recommendations to either refer a patient for further examination or not. Yet, the performance of current systems still typically makes some level of human grader verification necessary [3], which could be guided by an useful explanation of the AI system's decision. Also, clinical implementation would benefit from clinicians being able to understand the rationale behind the recommendation of the algorithm [5–7].

Typically, an AI system's decision are explained with heatmaps obtained post-hoc using gradient-based approaches [8–10]. However, such explanations are not trustworthy, as the produced heatmaps do not reflect the actual decision-making process of the model, and are prone to spurious correlations [11]. Therefore, their results cannot be easily integrated into the clinical decision-making process [7,12].

We address this issue and validate an inherently interpretable deep learning architecture for providing clinical decision support for screening for early DR in a retrospective reader study. Our approach uses a deep learning architecture called sparse BagNets [13,14], which explicitly models the local evidence for the presence of DR as part of its network architecture (Fig 1B). Most studies so far have considered the task of screening for moderate non-proliferative DR or more advanced stages [1], although even mild non-proliferative diabetic retinopathy (NPDR) is recommended for close monitoring and careful control of hyperglycemia [15,16]. We reasoned that the benefit of AI-based explanations and decision support would be most clearly visible for this challenging diagnostic task. Trained on a large

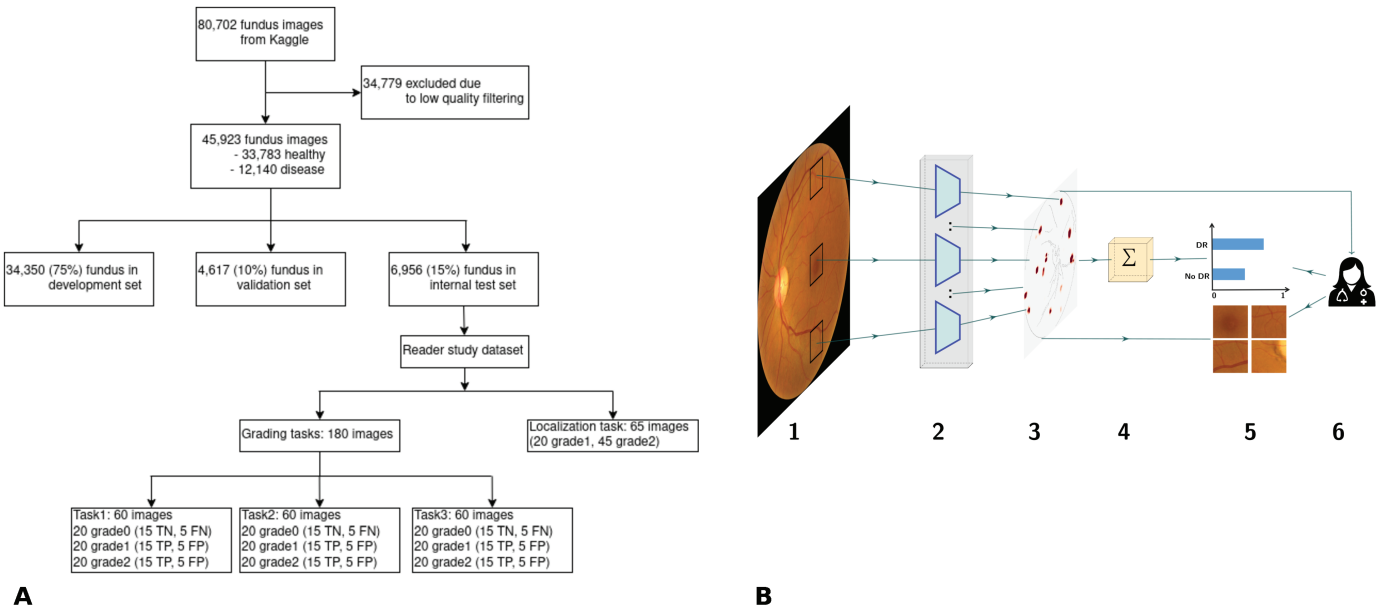

Fig 1. Overview of the development data and proposed inherently interpretable deep learning framework evaluated in this study. (A) Summary of the development dataset used to build the model, as well as the data used in the retrospective reader study. (B) Sparse BagNet architecture. 1. As a preliminary step, the retinal fundus image is implicitly split into many overlapping small patches of size 33 × 33. 2. All patches are fed to the model backbone, which processes them in parallel. 3. The BagNet backbone generates a heatmap that depicts the local disease evidence of individual patches. 4. The values of the heatmap are averaged and used as the final logit for classification. 5./6. The logits are fed into a softmax function which provides the probability distribution of the output, and then patches of suspect regions based on the heatmaps can be requested and viewed by a clinician to understand the classification results.

https://doi.org/10.1371/journal.pdig.0000831.g001

publicly available dataset, our model shows high specificity and sufficient sensitivity in detecting mild DR across a large array of datasets. Importantly, we show that the obtained class evidence maps highlight clinically relevant lesions such as microaneurysms or hemorrhages with high precision, making them useful for verifying the AI system's decisions. Finally, we show that the system can be effectively used to guide clinical decision-making, leading to 17.5% improvement in diagnostic accuracy for mild DR and overall about ≈ 25% improvement in screening time.

Methods

Dataset description and data preparation

We used eleven publicly available retinal image datasets, consisting of color fundus images from various sources, to develop and evaluate an inherently interpretable deep learning model for early DR detection (Table 1). For all datasets, fundus images had assigned reference grades based on the International Clinical Diabetic Retinopathy classification scale [17], which provides a grading scheme ranging from 0 (no DR), 1 (mild NPDR), 2 (moderate NPDR), 3 (severe NPDR) to 4 (proliferative DR) according to DR severity. As our goal was to develop an AI system for early DR screening, we combined class level {0} vs {1,2,3,4}. At stage 1, DR is in most cases asymptomatic, and challenging to detect even for experienced ophthalmologists. As all fundus datasets were fully anonymous, no approval from an Ethics Board was needed for this part of the study.

Table 1. Summary of the internal and external validation datasets used to evaluate the models. "Origin" refers to the country where the data was collected. "Lesion" refers to the number of images in the dataset with lesion annotations. The Kaggle dataset (first row, shaded in gray) is the internal dataset used to evaluate the model, while the other datasets were used for external validation to assess the generalization properties of the trained model.

Dataset	Origin	Number of images			Lesion
		All	Healthy	DR	
Kaggle [18]	USA	6,956	5,118	1,838	65
IDRiD [19]	India	512	168	348	81
E-Ophtha [20]	France	434	260	174	174
FGA-DR [21]	UAE	1,841	101	1,740	1,740
DIARETDB1 [22]	Finland	89	05	84	84
DDR [23]	China	12,513	6,265	6,248	755
DR2 [24]	Brazil	445	300	145	-
APTOS [25]	India	3,662	1,805	1,857	-
FCM-UNA [26]	Paraguay	757	187	570	-
Messidor-1 [27]	France	1,200	546	654	-
Messidor-2 [27,28]	France	1,744	1,017	727	-

https://doi.org/10.1371/journal.pdig.0000831.t001

Development dataset.

The dataset used to develop the inherently interpretable deep learning model was obtained from the Kaggle Diabetic Retinopathy challenge [18] which initially contained records of 44,351 subjects with 88,702 retinal fundus images from both eyes (Fig 1A). This dataset was originally provided by EyePacs Inc., a diabetes screening program in California. A comparable dataset also obtained from EyePacs Inc. included ethnicity information and contained about 70% images from patients with Latin American ethnicity [29]. We automatically quality filtered the fundus images using an ensemble of 10 EfficientNets models [30] trained on the DeepDRiD dataset [31]. This model achieved a quality filtering accuracy of 87.5% [32]. After quality filtering, we retained 45,923 images from 28,984 subjects for training, with 73% of images in the healthy class and 27% in the DR class. The dataset was split into training, validation, and test folds with 75%, 10%, and 15% of images, respectively, making sure that all images from the same subject were allocated to the same fold. The training fold was used for model fitting, the validation fold for model selection and hyperparameter tuning, and the test fold for internal evaluation.

To evaluate the explanations provided by the explainable sparse BagNet model, three ophthalmologists (authors AR, LaK, and NS with 5, 9, and 14 years of experience respectively) marked the location of DR-related lesions on 65 randomly selected fundus images from the test set (20 grade 1 and 45 grade 2) using a custom-written annotation browser interface (S1 Fig) based on the Python web framework Django, version 4.2.1, with a secure PostgreSQL database, version 15.3, and a Javascript front-end (available at https://github.com/berenslab/retimgtools/releases/tag/v1.1.0). Annotators were asked to mark "Microaneurysms (MA)", "Hemorrhages (HE)", "Exudates (EX)", "Soft Exudates (SE)" or "Other" for lesions visible on the fundus image. We combined the annotations of all graders into a consensus annotation for each image (S1 Table). We also assessed the consistency between ophthalmologists' annotations by calculating the dice between their annotations, showing that annotating DR-related lesions exhaustively is a challenging task (S2 Table).

External datasets.

Additional fundus data sets were obtained from various sources (Table 1) and were used for external evaluation of the model to assess the generalization performance. In addition to

reference DR grades, some of these external datasets [19–23] contained pixel-wise annotations for disease-related lesions. We used these additional annotations to evaluate the performance of the interpretable deep-learning model at localizing DR-related lesions.

Preprocessing.

Raw fundus images were preprocessed by cropping them to a square size of 512 x 512 pixels using a circle fitting method [33]. Then, image intensities were normalized by the mean and standard deviation of the training set. We applied this preprocessing procedure to all the fundus images from all datasets with the same parameters.

Inherently interpretable deep learning model for Diabetic Retinopathy detection

Architecture.

We trained and evaluated an inherently interpretable deep convolutional neural network (sparse BagNet [13,14]) for early DR detection. The sparse BagNet is an implicitly patch-based model based on bag-of-local features and aggregates local evidence from interpretable heatmaps to make predictions (Fig 1B). It takes a two-dimensional fundus image as input (Fig 1B.1) and outputs a binary prediction, which indicates the absence or presence of DR, together with the confidence as the probability score.

In contrast to other deep learning models, the sparse BagNet architecture is designed to be inherently interpretable, as the input image is implicitly split into many small, overlapping patches (size q = 33x33 pixels corresponding to the size of the model's effective receptive field with stride s = 8; Fig 1B.1), which are independently processed in parallel (Fig 1B.2) to compute the local evidence for the presence of DR. The patchwise predicted local evidence values are combined into a single class evidence map corresponding to a downsampled version of the input image (Fig 1B.3), which then is aggregated using average pooling and passed through a softmax function (Fig 1B.4) to output the probability distribution of DR (Fig 1B.5). Crucially, we employ a ℓ_1-penalty on the local evidence to encourage a sparse class evidence map.

After inference, the model can support screening not only with the final prediction but also with the class evidence map (Fig 1B.3) highlighting the contribution of small local regions to the final prediction. To this end, the evidence map is upsampled to the full image resolution and overlaid on the input image. In contrast to post-hoc gradient-based methods [11], the class evidence map provided by the sparse BagNet is a transparent part of the actual decision-making process and faithfully captures the local evidence. We supplement the class evidence map by extracting patches from regions with high DR evidence (Fig 1B.5).

Training procedure.

We trained the model on the training set by minimising the following loss function including the ℓ_1-penalty:

$$L\big((\mathbf{X},\theta),\mathbf{y}\big) = CE\big(f(\mathbf{X},\theta),\mathbf{y}\big) + \lambda \sum_{i,j,c} |\mathbf{A}_c^{ij}|. \tag{1}$$

Here, $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ denotes the input image with H,W,C being height, width, and the number of channels, CE is the cross-entropy, $\mathbf{y}$ are the reference class labels, f is the model with parameters θ, and $\mathbf{A}_c$ denotes the evidence map of class c. The sparsity of the evidence maps depends on the hyperparameter λ.

We initialized the model with weights pre-trained on ImageNet and then retrained and optimized for accuracy on the Kaggle DR dataset for 100 epochs. We used the stochastic gradient descent optimizer with an initial learning rate of 10^{-3}, and a clipped cosine learning rate scheduler with a minimum value set to 10^{-4}. We performed data augmentation during training by applying random cropping, flipping, color jitter, translation, and rotation following [34]. The sparsity hyperparameter λ was chosen based on the classification accuracy on the validation set (S2 Fig).

Baseline model and post-hoc interpretability

For comparison, we trained a standard black-box ResNet-50 [35] for early onset DR detection using the same training procedure as described above. We evaluated the classical interpretability techniques Integrated Gradients and Guided Backpropagation due to their high performance in identifying clinically validated DR lesions [36].

Clinical user study for AI-based decision support

Study dataset.

The user study was designed to evaluate the usefulness of the explanations provided by the inherently interpretable deep learning model in clinical practice. The dataset for each grading task (see below) consisted of 60 fundus images from the internal test set, where 20 images were sampled from grade 0, grade 1, and grade 2 respectively. For each grade, 15 images were correctly classified by the network and 5 falsely, making this a challenging screening task for clinicians. Thus, the fraction of images with DR in the user study was 66% and the deep learning model achieved an accuracy of 75% by design. Image grading was based solely on the fundus image and AI support, but no additional clinical data were provided.

Study design.

Six trained ophthalmologists with a median clinical experience of 9 years (4–17 years) participated in the reader study (including authors LaK, AR, and NS). We did not perform a formal power calculation. The study consisted of three tasks: In task 1 (referred to as "H"), participants were asked to grade fundus images without AI support (S3 Fig). In task 2 ("H+AI"), participants were additionally provided with the class predicted by the deep learning model and its confidence (S4 Fig). Finally, in task 3 ("H+XAI"), participants were additionally shown model explanations in the form of up to 12 bounding boxes around the regions from the class evidence map with the highest evidence, with bounding boxes matching the effective receptive field size and depicting the local image patches that contribute most to the global class evidence (S5 Fig).

For the three grading tasks, readers were instructed to classify each fundus image into two classes ("No DR" and "DR"). They were told to classify an image as "DR" even if they thought it only contained signs of mild non-proliferative DR (grade 1). None of the readers had access to the true labels. For task 3, readers were told that some bounding box explanations may contain healthy regions, as the algorithm also generated bounding boxes for healthy images erroneously classified as DR by the sparse BagNet model. In addition to the assigned class, we recorded the time it took for the reader to grade each image and asked them to rate their confidence on a scale from 1 to 5. Ethical approval for the study was obtained from the Ethics Committee at the University Hospital Tübingen (Ref No. 249/2023BO2).

A custom-written browser interface based on the Python web framework Django (version 4.2.1) with a secure PostgreSQL database (version 15.3) and a JavaScript front-end

was used to carry out the study (S3 Fig, S4 Fig, S5 Fig). The tool showed the fundus image, and response options and provided a digital magnifier to enlarge small image regions.

Evaluation criteria and statistical analysis

Criteria for evaluating the performance of the inherently interpretable deep learning model were specified before the start of the study based on previous work [13]. We evaluated three aspects of the model's quality:

1. DR screening performance compared to a regular deep learning model, within and across datasets.
2. The quality of the class evidence maps and derived bounding boxes in terms of lesion localization.
3. The usefulness of the inherently interpretable deep-learning model and the derived bounding boxes for decision support.

DR screening performance.

The primary measure of DR screening performance was the accuracy of the model for early DR detection using the reference labels. Additionally, we evaluated the area under the receiver-operating curve (AUC), sensitivity, specificity, and precision. All measures were computed on the internal test set as well as on the ten external datasets (Table 1). The model was not retrained or fine-tuned before assessment on the external datasets. All measures were computed using the scikit-learn package (v 1.0.2) and confidence intervals were computed using a bootstrap procedure with 1000 unstratified resamples [37].

Quality of class evidence maps.

To measure the quality of the class evidence maps and the derived bounding boxes for lesion localization, we calculated the proportion of highlighted regions (regions within the bounding box) that contained annotated lesions ("localization precision"). To this end, we used the annotations collected for this study on 65 images from the test set, as well as those external datasets containing pixel-level annotations (Table 1). We did not evaluate the fraction of lesions detected by our model ("recall"), as we did not train the model for lesion detection, and diagnostic support does not require an exhaustive detection of all lesions.

Statistical analysis of decision support.

We measured the performance of the readers in our clinical user study as the accuracy of the reader's decision with respect to the reference labels. To assess the effect of the task and DR reference grade statistically, we fit the responses with a generalized linear model (R, function *glm*, v 4.0.3) with predictor *task* or with predictors *task* and *DR grade* including interactions. If we found significant predictors at the $\alpha = 0.05$ level, we computed the marginal means and 95%-confidence intervals (package *emmeans*, v 1.5.3) as well as the respective contrasts between conditions for post-hoc testing. Tukey's method was used for correcting for multiple comparisons. We used the same procedure for analyzing the measured grading time and the reported confidence, but used a linear model (function *lm*) instead.

Role of the funding source

The funders of this work had no role in the study design, collection, analysis, and interpretation of data, the writing of the report, nor in the decision to submit the paper for publication.

Results

We trained and evaluated an inherently interpretable deep learning model ("sparse BagNet") for early DR screening (Fig 1B). We first evaluated screening performance for early DR against the state-of-the-art non-interpretable black-box model ("ResNet50") on the internal test set of the development dataset and on a large number of additional datasets (see Table 2). The sparse BagNet performed well and was comparable to the state-of-the-art model on the internal test set (accuracy: 0.906, 95% CI [0.900–0.913]; AUC: 0.904 [0.894–0.913]; sensitivity: 0.709 [0.688–0.729]; specificity: 0.977 [0.973–0.981]; precision: 0.918 [0.903–0.932]) and generalized well to a number of external datasets (Table 2).

The key advantage of our inherently interpretable model is that the local disease evidence is explicitly represented in a class evidence map (Fig 1B.3 and Fig 2B). During training, the class evidence map is encouraged to be sparse, such that the final loss function balances prediction accuracy and an interpretable map. For the model studied above, the regularization parameter trading-off accuracy and sparseness was heuristically chosen such that sparseness was encouraged at a minimal loss of accuracy (S2 Fig). At each location in the class activation map, the color indicates the model output for an individual image patch. We detected the regions with the highest evidence and placed bounding boxes corresponding to the patch size around these points (Fig 2A).

Table 2. Summary of the classification performance with confidence intervals (CIs) computed at 95% using bootstrapping (n=1000). "AUC" refer to the receiver-operating curve. "Loc Bag" and "Loc GBP" respectively refer to the localization precision of the sparse BagNet and Guided Backpropagation on ResNet-50 at localizing lesions from annotated images. For each dataset, the first row shows the performance of the interpretable sparse BagNet model, while the second row shows the performance of the baseline black-box ResNet-50 model. The Kaggle dataset (first row) is the internal dataset used to train and evaluate the model, while the other datasets were used for external validation to assess the generalization properties of the trained model. The low classification performance on the FCM-UNA and FGA-DR datasets can be explained by the relatively low quality of most images in the FCM-UNA dataset and the large intensity variation of the FGA-DR dataset (S6 Fig). The low localization precision (0.664) on the E-Ophtha dataset is likely due to annotations only being provided for "Microaneurysms" and "Exudate" lesions, while the images could contain other DR-related lesions.

Dataset		Accuracy	AUC	Sensitivity	Specificity	Precision	Loc_{Bag}	Loc_{GBP}
Kaggle	Bag.	.906 (.900 - .913)	.904 (.894 - .913)	.709 (.688 - .729)	.977 (.973 - .981)	.918 (.903 - .932)	.941	-
	Res.	.914 (.907 - .921)	.935 (.927 - .943)	.765 (.745 - .784)	.967 (.962 - .972)	.894 (.878 - .908)	-	.656
		.891 (.864 - .917)	.879 (.838 - .913)	.951 (.927 - .972)	.768 (.699 - .828)	.895 (.861 - .925)	.804	-
IDRiD		.882 (.851 - .909)	.864 (.822 - .902)	.963 (.942 - .981)	.714 (.639 - .781)	.875 (.84 - .908)	-	.140
		.903 (.864 - .917)	.944 (.838 - .913)	.920 (.927 - .972)	.892 (.699 - .828)	.851 (.861 - .925)	.656	-
E-Ophtha		.933 (.851 - .909)	.972 (.822 - .902)	.966 (.942 - .981)	.912 (.639 - .781)	.880 (.840 - .908)	-	.030
		.799 (.781 - .819)	.789 (.752 - .823)	.811 (.793 - .830)	.594 (.500 - .687)	.972 (.963 - .980)	.872	-
FGA-DR		.763 (.743 - .781)	.816 (.768 - .858)	.764 (.743 - .783)	.743 (.653 - .819)	.981 (.973 - .987)	-	.336
		.831 (.753 - .899)	.931 (.870 - .981)	.821 (.733 - .898)	1	1	.881	-
DIARETDB1		.742 (.652 - .831)	.811 (.715 - .900)	.738 (.640 - .829)	.800 (.333 - 1.00)	.984 (.950 - 1.00)	-	.000
		.825 (.818 - .832)	.926 (.922 - .931)	.669 (.657 - .681)	.980 (.977 - .984	.971 (.966 - .976)	.965	-
DDR		.887 (.881 - .892)	.963 (.960 - .966)	.800 (.790 - .810)	.973 (.968 - .977)	.967 (.962 - .972)	-	.249
		.879 (.847 - .908)	.922 (.889 - .951)	.662 (.584 - .742)	.983 (.968 - .997)	.950 (.905 - .990)	-	
DR2		.876 (.845 - .906)	.866 (.825 - .905)	.669 (.591 - .742)	.977 (.959 - .993)	.933 (.884 - .975)	-	
		.973 (.968 - .979)	.995 (.992 - .996)	.982 (.975 - .987)	.965 (.956 - .973)	.966 (.958 - .974)	-	
APTOS		.949 (.942 - .956)	.972 (.965 - .978)	.942 (.931 - .952)	.956 (.946 - .965)	.956 (.947 - .966)	-	
		.773 (.744 - .802)	.936 (.918 - .952)	.702 (.664 - .738)	.989 (.972 - 1.00)	.995 (.987 - 1.00)	-	
FCM-UNA		.877 (.853 - .900)	.967 (.954 - .979)	.840 (.811 - .868)	.989(.971 - 1.00)	.996 (.989 - 1.00)	-	
		.889 (.871 - .907)	.943 (.929 - .955)	.832 (.804 - .859)	.958 (.939 - .974)	.959 (.941 - .975)	-	
Messidor-1		.893 (.876 - .909)	.954 (.942 - .965)	.852 (.823 - .878)	.943 (.923 - .963)	.947 (.928 - .964)	-	
		.829 (.812 - .847)	.876 (.859 - .894)	.750 (.719 - .785)	.886 (.865 - .906)	.825 (.794 - .853)	-	
Messidor-2		.851 (.835 - .869)	.925 (.912 - .938)	.794 (.763 - .823)	.893 (.875 - .913)	.841 (.815 - .868)	-	

https://doi.org/10.1371/journal.pdig.0000831.t002

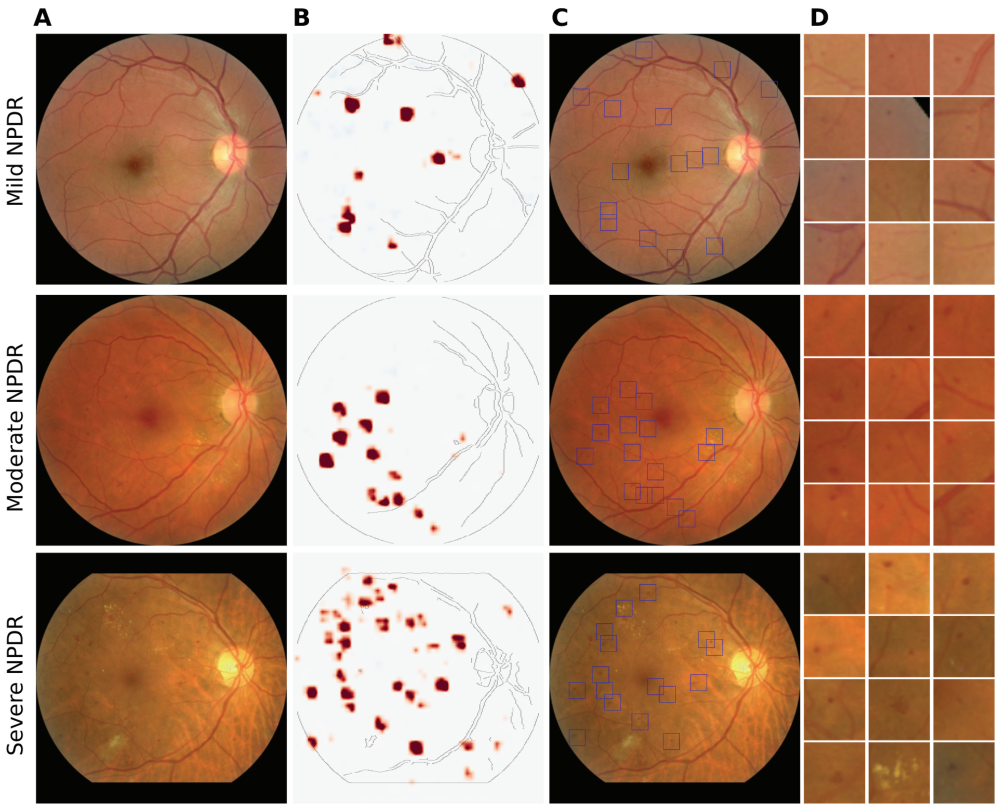

Fig 2. Inherently interpretable deep learning framework highlights clinically relevant image regions. (A) Examples of retinal fundus images from different DR grades (top to bottom: mild NPDR, moderate NPDR and severe NPDR). (B) Class evidence map extracted from the inherently interpretable model without further processing. Red regions indicate evidence for the presence of at least mild DR. (C) Bounding boxes drawn around suspicious regions in the class evidence map. In some cases, the bounding boxes are placed in regions for which there is no visible evidence due to the scaling of the color map. Yet, these evidence values are also strictly positive. (D) Suspicious regions from (C) enlarged and sorted with decreasing evidence scores. Depending on the image grade, the suspicious regions contain various DR-related lesions such as microaneurysms, hemorrhages, or drusen.

https://doi.org/10.1371/journal.pdig.0000831.g002

Although the model was never trained with pixel-level annotations or supervision signals other than the image-level DR reference label, the highlighted regions typically contained DR-related lesions such as microaneurisms, drusen, or hemorrhage with high precision (Fig 3).

We quantitatively evaluated how well the class evidence maps provided information about the location of disease-related lesions using a subset of images from the test set of the development dataset (Fig 3) as well as external datasets with pixel-level annotations (Table 1). The class evidence maps precisely localized DR lesions, as most regions flagged as suspicious indeed contained annotated lesions (Table 2, last column). For the images from the development dataset, we obtained a precision of 0.960 (95% CI [0.941–0.976]), with minor differences between images with mild and moderate NPDR (0.783 vs. 0.970). Notably, our model generalized well to external test sets, with precision ranging from 0.656 to 0.965 (Table 2, last column).

We also evaluated suspicious regions extracted from images the algorithm falsely classified as DR with high confidence (>0.75). To this end, we showed two clinicians 30 images falsely classified as DR with bounding boxes (S8 Fig). Sometimes, these image patches showed

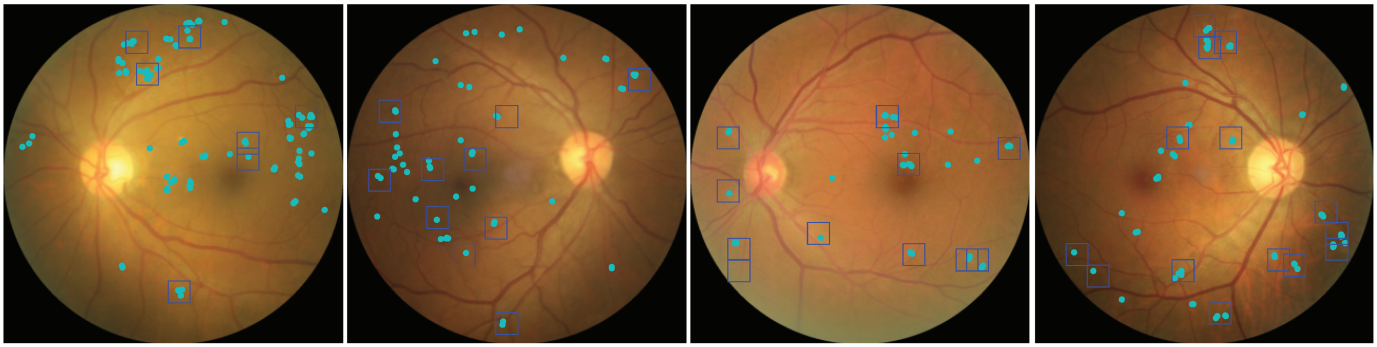

Fig 3. Extracted high evidence images patches contain DR-related lesions. Example fundus images with DR, with DR lesions identified by three clinicians (cyan). Bounding boxes (blue) were extracted from the class evidence maps based on regions of high evidence for DR. Note that all bounding boxes contain annotated lesions, but – as the number of bounding boxes per image was restricted to twelve – not all lesions are contained in bounding boxes.

https://doi.org/10.1371/journal.pdig.0000831.g003

unclear or ambiguous lesions unrelated to DR, but they typically contained anomalies related to DR such as microaneurysms or exudates, but not in a number or severity sufficient for clinical DR diagnosis (S8 Fig).

We next compared the localization performance of the inherently interpretable sparse BagNet to classic post-hoc methods such as Integrated Gradients [38] or Guided Backprop applied to the state-of-the-art model (Fig 4A–4C). These methods were chosen because they performed well in a clinical validation of post-hoc explainability techniques for DR [36]. We found that bounding boxes obtained from Guided Backprop or Integrated Gradients were much less precise in localizing DR-related lesions (0.941 vs. 0.656, Fig 4D, Table 2), especially for out-of-sample test datasets.

We then investigated whether our interpretable deep learning model could effectively aid clinicians in detecting DR via a retrospective reader study with six experienced ophthalmologists screening fundus images for the presence of early DR with various levels of AI assistance (see Methods). Without AI assistance (labeled "H") ophthalmologists reached a mean classification accuracy of 0.611 (95% CI [0.560–0.660]; Fig 5A). Their accuracy increased significantly to 0.758 ([0.711–0.800], $p = 0.0001$, post-hoc test with Tukey's correction for multiple comparisons, see Methods) when they had access to the deep learning model's prediction and confidence ("H+AI"). They achieved similar performance with additional access to AI explanations in the form of bounding boxes around suspicious regions extracted from the class evidence maps ("H+XAI") at an accuracy of 0.786 [0.741–0.825].

We studied ophthalmologists' performance in screening for DR in fundus images of different disease grades in more detail (Fig 5B). Without AI support, detecting images with mild DR (grade 1) was the most challenging with comparably low performance, which improved with AI support. For healthy images, screening performance improved significantly with any form of AI decision support (H: 0.567, [0.477–0.652]; H+AI: 0.842, [0.765–0.897]; H+XAI: 0.817, [0.737–0.876]; H vs. H+AI: $p < 0.0001$; H vs. H+XAI: $p = 0.0001$; H+AI vs. H+XAI: $p = 0.8645$), while for images with mild DR, we observed that screening only improved significantly for AI support with explanations (H: 0.483, [0.395–0.572]; H+AI: 0.617, [0.527–0.699]; H+XAI: 0.733, [0.647–0.805]; H vs. H+AI: $p = 0.0962$; H vs. H+XAI: $p = 0.0003$; H+AI vs. H+XAI: $p = 0.1326$). For images with moderate DR, AI support had no significant effect on screening performance. Taken together, this provides evidence that giving ophthalmologists access to AI support led to superior DR screening performance, with explanations based on the sparse BagNet model being most effective for difficult diagnostic decisions.

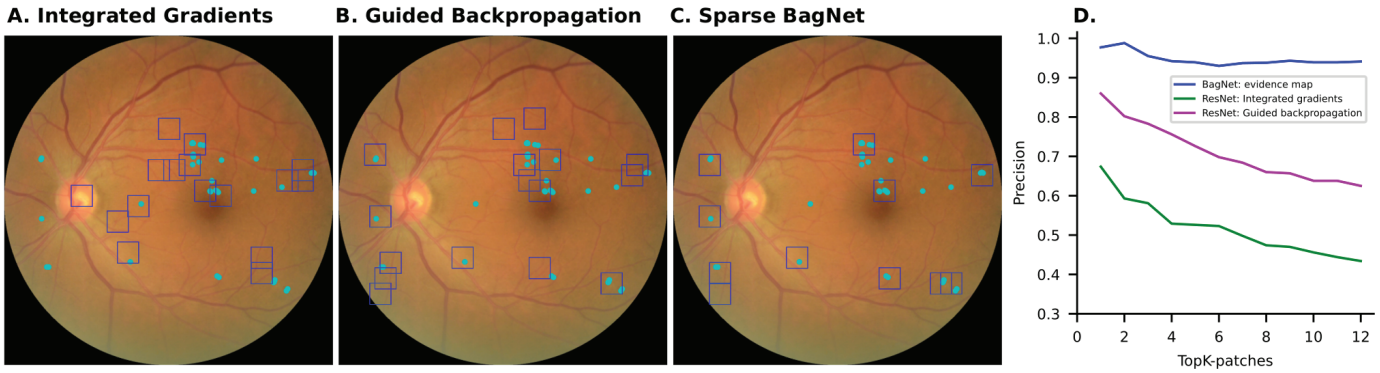

Fig 4. Inherently interpretable deep learning framework highlights lesions more precisely than post-hoc techniques applied to a standard DNN. (A) Suspicious regions (blue) marked with bounding boxes extracted from the heatmap obtained with Integrated gradients from the standard DNN. Clinically relevant DR lesions are marked in cyan. (B) As in (A) extracted from the heatmap obtained with Guided backpropagation. (C) For comparison, suspicious regions were obtained from the SparseBagNet. (D) Systematic comparison of localization precision for clinically annotated DR lesions as a function of the number of considered patches.

https://doi.org/10.1371/journal.pdig.0000831.g004

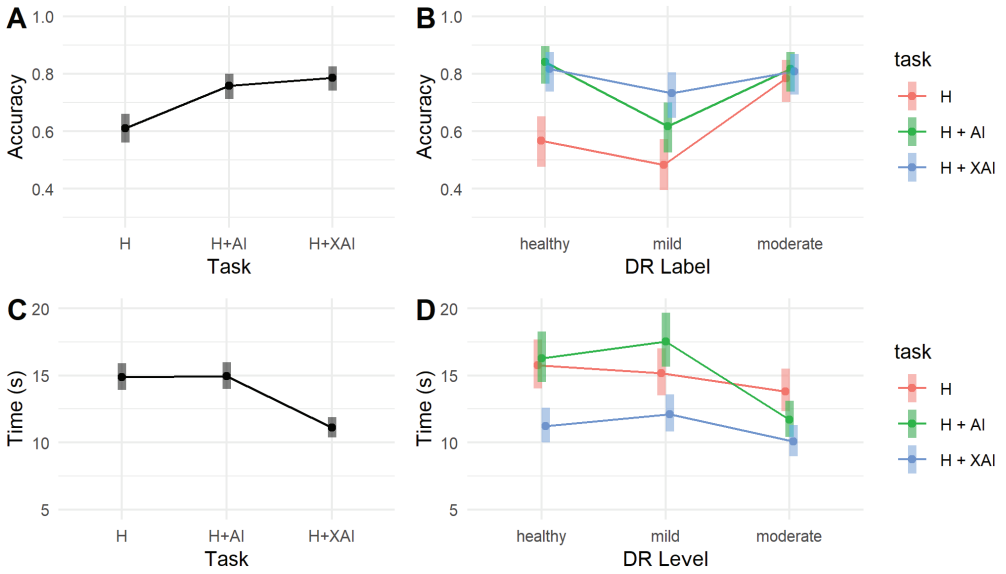

Fig 5. Providing AI-based clinical decision support based on the inherently interpretable deep learning model improves DR screening. (A) Screening accuracy with different levels of AI assistance. Six ophthalmologists graded fundus images without AI assistance ("H"), with access to the AI prediction ("H+AI"), and with additional access to AI explanations ("H+XAI"). AI assistance improved screening accuracy, but access to AI explanations had only a small additional effect. (B) Screening accuracy for DR on fundus images of different disease grades. For healthy images, accuracy improved significantly with any form of AI decision support ("H+AI" or "H+XAI"), while for images with mild DR, screening improved significantly for AI support with explanation ("H+XAI"). For images with moderate DR, AI support had no significant effect on screening performance. (C) Screening time in screening DR with different levels of AI assistance. The decision time is significantly reduced with AI support ("H+XAI") with explanation compared to the other tasks ("H", and "H+AI"). (D) Screening time in screening for DR on fundus images of different disease grades. Screening time reduces at all disease stages with a significant effect of AI decision support with explanation for healthy images ("grade 0"), mild DR ("grade 1"), and moderate DR ("grade 2").

https://doi.org/10.1371/journal.pdig.0000831.g005

We next studied whether AI decision support would not only allow ophthalmologists to make more accurate screening decisions but also reach their decisions faster. We found that

the decision time was significantly reduced when providing ophthalmologists AI support with explanations compared to both other tasks (Fig 5A, H: 15.2 s [14.1-16.4]; H+AI: 15.9 s [14.7-17.1]; H+XAI: 11.7 s [10.8-12.6]; H vs. H+AI: $p = 0.7435$; H vs. H+XAI: $p < 0.0001$; H+AI vs. H+XAI: $p < 0.0001$). This reduction was present at all disease stages, with a significant effect of AI decision support with explanations for healthy images (Fig 5A; H: 15.8 s [14.1-17.7]; H+AI: 16.3 s [14.5-18.3]; H+XAI: 11.2 s [10.0-12.6], H vs. H+AI: $p = 0.9153$; H vs. H+XAI: $p < 0.0001$; H+AI vs. H+XAI: $p < 0.0001$), mild DR (H: 15.2 s [13.5-17.0]; H+AI: 17.5 s [15.6-19.7]; H+XAI: 12.1 s [10.8-13.6], H vs. H+AI: $p = 0.1843$; H vs. H+XAI: $p = 0.180$; H+AI vs. H+XAI: $p < 0.0001$), as well as moderate DR (H: 13.8 s [12.3-15.5]; H+AI: 11.7 s [10.4-13.1]; H+XAI: 10.1 s [9.0-11.3]; H vs. H+AI: $p = 0.1058$; H vs. H+XAI: $p = 0.004$; H+AI vs. H+XAI: $p = 0.1724$). In summary, this indicates that decision support with accurate explanations provided by the sparse BagNet model could reduce screening times across all disease levels.

We also analyzed whether AI decision support would change the confidence with which the ophthalmologists could grade the images, but did not find a significant effect of AI support (H: 3.8 [3.7-3.9]; H+AI: 3.7 [3.6-3.9]; H+XAI: 3.6 [3.5-3.7], H vs. H+AI: $p = 0.6806$; H vs. H+XAI: $p = 0.0543$; H+AI vs. H+XAI: $p = 0.3023$). We conclude that self-reported confidence may not be a reliable measure of grader uncertainty compared to recorded decision time.

We finally analyzed whether the positive effect on accuracy was dependent on whether the deep learning model had classified the image correctly or not, as AI support has been reported to be detrimental in case of model errors [39]. In line with the results above, we found that screening performance and decision time significantly improved for cases in which the deep learning model had made a correct decision (S5 Fig; accuracy, H vs. H+AI: $p<0.0001$; H vs. H+XAI: $p<0.0001$; H+AI vs. H+XAI: $p<0.0001$; time, H vs. H+AI: $p = 0.8178$; H vs. H+XAI: $p<0.0001$; H+AI vs. H+XAI: $p<0.0001$). For cases in which the model had made an incorrect decision, we neither detected positive nor negative effects on accuracy (H vs. H+AI: $p<0.3216$; H vs. H+XAI: $p = 0.4953$; H+AI vs. H+XAI: $p = 0.9480$) and slightly positive effects on decision time (H vs. H+AI: $p = 0.4557$; H vs. H+XAI: $p = 0.0941$; H+AI vs. H+XAI: $p = 0.0031$) meaning that the decision time was still smaller despite the wrong prediction of the model.

Discussion

In this study, we trained and evaluated an inherently interpretable deep-learning model for early diabetic retinopathy detection. This is a challenging task even for experienced ophthalmologists. Our model achieved a classification performance comparable to the black-box baseline model in the internal test set and on ten publicly available external datasets. While the training dataset contained a large fraction of images from patients of Latin American ethnicity, the external datasets were acquired in diverse world regions and different devices, thus that our model showed a good generalization across different ethnic groups and patient populations. While some of these datasets also contained patients of African ancestry, none of the datasets were acquired on the African continent.

In addition to a binary diagnostic decision that is commonly communicated in DR screening settings, our model provides explanations via interpretable evidence maps, which highlight regions of the image used by the network in making its decisions. We found that the inherently interpretable framework precisely located disease-related lesions in the image,

more so than post-hoc techniques applied to a state-of-the-art DNN, in particular for out-of-sample test datasets. Even in case of incorrect model predictions according to the reference labels, model explanations proofed to be useful and highlighted suspicious regions.

In a retrospective reader study, we found that highlighting these regions during grading helped ophthalmologists improve their grading performance, especially for difficult cases, while reducing their decision time. This indicates that current paradigms used in AI-based screening scenarios may benefit from including explanations for easier human verification and enhanced trust in the algorithms decision [3,5]. Our study further showed that the errors of the AI model did not negatively affect decision-making by ophthalmologists, in contrast to earlier human-AI studies on clinical decision support [39,40]. A limitation of our model is that it was trained on a dataset from North America, and may need to be fine-tuned on data from the intended target population, although its generalization results on ten additional datasets were promising.

As the potential of AI for medical image analysis has become evident [41,42], such systems have reached performance close to, or even superior to, those of clinical experts in a variety of tasks [43]. More recently, the focus has shifted towards AI systems assisting clinicians in making better decisions [39]. In this setting, clinicians need to understand how decisions are formed by the AI model, such that transparency and interpretability of medical AI systems have become important aspects [7,11,12,44]. In agreement, the need for trustworthy and transparent AI systems and effective human/AI collaboration has been identified in standardized guidelines to facilitate their adoption in clinical practice [44]. While this generally poses challenges in balancing high performance and interpretability [45], our study has shown that inherent interpretability can be achieved without significant performance trade-offs if the inductive biases of the interpretable model are met – in our case, as early DR causes only very localized lesions in the retina. Other inherently interpretable models include prototype-based networks [46], which are difficult to use for diseases with many small, distributed lesions, for which the training procedure is more complex and for which interpretability is not straightforward [47].

In a clinical setting, such an inherently interpretable model could assist clinicians in mitigating the challenge of early and accurate diagnosis of presymptomatic diseases, such as diabetic retinopathy detection. Given several approved AI systems for DR screening, clinical implementation could be comparatively straightforward. The trained model can efficiently generate predictions, on a time scale not impeding on clinical practice ($\ll 1s$/image), requires relatively little memory ($\sim 350mb$), and does not require additional models to run to create explanations. Such explanations could be added to existing reporting templates in commercial AI systems, allowing screeners to quickly ascertain the plausibility of the models prediction. In this setting, real-world prospective studies could be conducted to test the impact of the explanations obtained from our model on screening quality and speed, in particular for patient with beginning DR.

One limitation of our model is that it may not provide good explanations if its inductive bias is not matched to the disease, e.g. when lesions cover large parts of the retina as in more advanced DR grades [13]. Future applications also include time-to-progression prediction for diseases like DR [48] through interpretable-by-design deep survival models [49].

Supporting information

S1 Fig. Web interface for the annotation task. A fundus image is shown and based on it, the annotator is asked to annotate lesions related to Diabetic Retinopathy. By moving the mouse over a region of the image, an enlarged version of that region is displayed. All images are from

patients with DR of grade 1 ("mild DR") or 2 ("moderate DR"). Each lesion is marked by selecting the type (Microaneurysms: MA, hemorrhages: HE, exudates: EX, soft exudate: SE, artifact, or any other lesions) and clicking on the image location.
(TIF)

S1 Table. **Summary of model performance on localizing DR-related lesions from graders' annotations** The precision of the model on each clinician annotation is calculated as the proportion of bounding boxes from regions highlighted on heatmaps containing lesions annotated by a grader. The random precision is obtained by drawing 20 random bounding boxes over each annotated image, excluding those falling in regions containing more than 10% black pixels. The union "∪" gives the precision of the model with the combined clinicians' annotation masks, while the intersection "∩" gives the precision of the model with reference annotation masks obtained as the intersections of clinicians' annotation over each image.
(PDF)

S2 Table. **Inter-grader performance on 65 fundus images from the internal Kaggle test set annotated by three ophthalmologists** "Grader X - Grader Y" refers to the dice score between grader X and grader Y. The Dice score is calculated for each pair of graders as the overlap between their annotation using a patch size of 33×33 pixels corresponding to the receptive field of the model and considering different strides (s = 8, 32 for overlapping patches and s=33 for non-overlapping patches). "Grader X - Grader Y ∪ Grader Z" refers to the dice score between grader X, Y, and Z while "Grader Y ∪ Grader Z" is the union between grader Y and Z, and "Grader Y ∩ Grader Z" is the intersection between grader Y and Z.
(PDF)

S2 Fig. **Comparison of the sparse BagNet performance with different regularization values on the validation dataset** The regularization coefficient λ affects the classification performance (accuracy and AUC) of the model. The red points indicate the selected value, which is a compromise between sparsity and both accuracy and AUC. It also defines the trade-off between the model's interpretability and classification performance.
(TIF)

S3 Fig. **Web interface for the grading task without AI support ("H")** A fundus image is shown and based on it, the grader is asked to decide whether the corresponding patient has Diabetic Retinopathy (DR) of any severity, including mild DR. In addition, the grader is asked to rate the confidence of his/her decision on a scale from 1 (least confident) to 5 (most confident). By moving the mouse over a region of the image, an enlarged version of that region is displayed. The time taken to reach each decision (grading and confidence) is recorded.
(TIF)

S4 Fig. **Web interface for the grading task with AI support ("H + AI")** A fundus image is shown with the model's prediction and its confidence level (from 0% to 100 %, with 100% being the highest confidence score). Based on this, the grader is asked to decide whether the corresponding patient has Diabetic Retinopathy (DR) of any severity, including mild DR. In addition, the grader is asked to rate the confidence of his/her decision on a scale from 1 (least confident) to 5 (most confident). By moving the mouse over a region of the image, an enlarged version of that region is displayed. The time taken to reach each decision (grading and confidence) is recorded.
(TIF)

S5 Fig. **Web interface for the grading task with AI support and explanations ("H + XAI").** A fundus image is shown with the model's prediction, its confidence level (from 0% to 100 %,

with 100% being the highest confidence score), and explanation in the form of blue bounding boxes around the regions for which the AI model believes that they contain signs of DR. Based on this, the grader is asked to decide whether the corresponding patient has Diabetic Retinopathy (DR) of any severity, including mild DR. In addition, the grader is asked to rate the confidence of his/her decision on a scale from 1 (least confident) to 5 (most confident). By moving the mouse over a region of the image, an enlarged version of that region is displayed. The time taken to reach each decision (grading and confidence) is recorded.
(TIF)

S6 Fig. Examples of fundus images from each dataset.
(TIF)

S7 Fig. Heatmap with combined clinicians' annotations of four examples of fundus cases with DR. For each example, the left side shows the heatmap with clinicians' annotations and bounding boxes around the regions of positive activation, while the right side shows the fundus image with clinicians' annotations and bounding boxes around the regions of positive activation.
(TIF)

S8 Fig. Examples of high-confidence false positives analyzed by two clinicians. On the left side of each example, the image displays bounding boxes highlighting regions with positive activation. On the right side, the suspicious regions from the left are enlarged and arranged in descending order of evidence scores. (A) A false-positive image where the clinicians interpreted the suspicious regions as "vitreous opacities" and "uveitis vitreous cells," respectively. (B) A false-positive image where one clinician identified the suspicious regions as "synchisis scintillans," while the other suggested the patient may have recently received an intravitreal steroid injection. (C) A false-positive image where both clinicians identified the suspicious regions as "microaneurysms" possibly associated with bleeding. (D) A false-positive image where both clinicians recognized the suspicious regions as "microaneurysms" and "hard exudates". **(E, F)** False-positive images where one clinician classified the image as DR while the other classifies it as no DR, citing the presence of only a single microaneurysm lesion in the suspicious regions.
(TIF)

S9 Fig. Analysis of errors of the AI model on accuracy and decision times for different tasks during the retrospective reader study. (a) For all tasks, ophthalmologists' accuracy is higher when the deep learning model makes the correct decision. For correct classifications, the AI assistance improves grading accuracy. For incorrect classification, it does not make it worse. **(b)** Ophthalmologists' decision time decreases overall when the deep learning model makes the correct decision. When the AI model is correct, the explanation decreases decision time significantly, while it does not increase the decision time for incorrect decisions.
(TIF)

Acknowledgments

We thank Sarah Müller, Pearse Keane and Murat Ayhan for discussion and Murat Ayhan quality filtering code.

Author contributions

Conceptualization: Kerol Djoumessi, Lisa M. Koch, Philipp Berens.

Data curation: Kerol Djoumessi, Ziwei Huang, Annekatrin Rickmann, Natalia Simon, Lisa M. Koch.

Formal analysis: Kerol Djoumessi.

Funding acquisition: Philipp Berens, Lisa M. Koch.

Investigation: Kerol Djoumessi, Laura Kühlewein, Lisa M. Koch, Philipp Berens.

Methodology: Kerol Djoumessi, Philipp Berens.

Project administration: Philipp Berens.

Software: Kerol Djoumessi, Ziwei Huang.

Supervision: Lisa M. Koch, Philipp Berens.

Validation: Laura Kühlewein, Annekatrin Rickmann, Natalia Simon.

Visualization: Kerol Djoumessi.

Writing – original draft: Kerol Djoumessi, Lisa M. Koch, Philipp Berens.

Writing – review & editing: Laura Kühlewein, Annekatrin Rickmann, Natalia Simon.

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):2402–10. https://doi.org/10.1001/jama.2016.17216 PMID: 27898976
2. Food US, Administration D. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices (SaMD) Action Plan; 2021. Available from: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices
3. Grzybowski A, Brona P, Lim G, Ruamviboonsuk P, Tan GSW, Abramoff M, et al. Artificial intelligence for diabetic retinopathy screening: a review. Eye (Lond). 2020;34(3):451–60. https://doi.org/10.1038/s41433-019-0566-0 PMID: 31488886
4. Ipp E, Liljenquist D, Bode B, Shah VN, Silverstein S, Regillo CD, et al. Pivotal evaluation of an artificial intelligence system for autonomous detection of referrable and vision-threatening diabetic retinopathy. JAMA Netw Open. 2021;4(11):e2134254. https://doi.org/10.1001/jamanetworkopen.2021.34254 PMID: 34779843
5. Grauslund J. Diabetic retinopathy screening in the emerging era of artificial intelligence. Diabetologia. 2022;65(9):1415–23. https://doi.org/10.1007/s00125-022-05727-0 PMID: 35639120
6. Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. J Med Ethics. 2020;46(3):205–11. https://doi.org/10.1136/medethics-2019-105586 PMID: 31748206
7. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206–15. https://doi.org/10.1038/s42256-019-0048-x PMID: 35603010
8. Chetoui M, Akhloufi MA. Explainable diabetic retinopathy using EfficientNET. Annu Int Conf IEEE Eng Med Biol Soc. 2020;2020:1966–9. https://doi.org/10.1109/EMBC44109.2020.9175664 PMID: 33018388
9. Alghamdi HS. Towards explainable deep neural networks for the automatic detection of diabetic retinopathy. Appl Sci. 2022;12(19):9435. https://doi.org/10.3390/app12199435
10. Gonzalez-Gonzalo C, Liefers B, van Ginneken B, Sanchez CI. Iterative augmentation of visual evidence for weakly-supervised lesion localization in deep interpretability frameworks: application to color fundus images. IEEE Trans Med Imaging. 2020;39(11):3499–511. https://doi.org/10.1109/TMI.2020.2994463 PMID: 32746093
11. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745–50. https://doi.org/10.1016/S2589-7500(21)00208-9 PMID: 34711379

12. Grote T, Berens P. How competitors become collaborators-Bridging the gap(s) between machine learning algorithms and clinicians. Bioethics. 2022;36(2):134–42. https://doi.org/10.1111/bioe.12957 PMID: 34599834
13. Kerol D, Ilanchezian I, Kühlewein L, Faber H, Baumgartner C, Bah B, et al. Sparse activations for interpretable disease grading. Med Imaging Deep Learn. 2023.
14. Brendel W, Bethge M. Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. 2019:1–10. https://doi.org/10.1109/ICLR.2019.00001
15. Solomon SD, Chew E, Duh EJ, Sobrin L, Sun JK, VanderBeek BL, et al. Diabetic retinopathy: a position statement by the american diabetes association. Diabetes Care. 2017;40(3):412–8. https://doi.org/10.2337/dc16-2641 PMID: 28223445
16. Vujosevic S, Aldington SJ, Silva P, Hernández C, Scanlon P, Peto T, et al. Screening for diabetic retinopathy: new perspectives and challenges. Lancet Diabetes Endocrinol. 2020;8(4):337–47. https://doi.org/10.1016/S2213-8587(19)30411-5 PMID: 32113513
17. Wilkinson CP, Ferris FL 3rd, Klein RE, Lee PP, Agardh CD, Davis M, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology. 2003;110(9):1677–82. https://doi.org/10.1016/S0161-6420(03)00475-5 PMID: 13129861
18. Dugas E, Jared J, Cukierski W. Diabetic Retinopathy Detection; 2015. Available from: https://kaggle.com/competitions/diabetic-retinopathy-detection
19. Porwal P, Pachade S, Kamble R, Kokare M, Deshmukh G, Sahasrabuddhe V, et al. Indian Diabetic Retinopathy Image Dataset (IDRiD): a database for diabetic retinopathy screening research. Data. 2018;3(3):25. https://doi.org/10.3390/data3030025
20. Decencière E, Cazuguel G, Zhang X, Thibault G, Klein J-C, Meyer F, et al. TeleOphta: machine learning and image processing methods for teleophthalmology. IRBM. 2013;34(2):196–203. https://doi.org/10.1016/j.irbm.2013.01.010
21. Zhou Y, Wang B, Huang L, Cui S, Shao L. A benchmark for studying diabetic retinopathy: segmentation, grading, and transferability. IEEE Trans Med Imaging. 2021;40(3):818–28. https://doi.org/10.1109/TMI.2020.3037771 PMID: 33180722
22. Kauppi T, Kalesnykiene V, Kamarainen JK, Lensu L, Sorri I, Raninen A, et al. The diaretdb1 diabetic retinopathy database and evaluation protocol. BMVC. 2007;1:10.
23. Li T, Gao Y, Wang K, Guo S, Liu H, Kang H. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Inf Sci. 2019;501:511–22.
24. Pires R, Jelinek HF, Wainer J, Valle E, Rocha A. Advancing bag-of-visual-words representations for lesion classification in retinal images. PLoS One. 2014;9(6):e96814. https://doi.org/10.1371/journal.pone.0096814 PMID: 24886780
25. Karthik SD Maggie. APTOS 2019 blindness detection; 2019. Available from: https://kaggle.com/competitions/aptos2019-blindness-detection
26. Benítez-VEC MI, Román J, Noguera J, García-Torres M, Ayala J. Dataset from fundus images for the study of diabetic retinopathy. Data Brief. 2021;36:107068. https://doi.org/10.1016/j.dib.2021.107068
27. Decencière E, Zhang X, Cazuguel G, Lay B, Cochener B, Trone C, et al. Feedback on a publicly distributed image database: the messidor database. Image Anal Stereol. 2014;33(3):231. https://doi.org/10.5566/ias.1155
28. Abràmoff MD, Folk JC, Han DP, Walker JD, Williams DF, Russell SR, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. JAMA Ophthalmol. 2013;131(3):351–7. https://doi.org/10.1001/jamaophthalmol.2013.1743 PMID: 23494039
29. Koch LM, Baumgartner CF, Berens P. Distribution shift detection for the postmarket surveillance of medical AI algorithms: a retrospective simulation study. NPJ Digit Med. 2024;7(1):120. https://doi.org/10.1038/s41746-024-01085-w PMID: 38724581
30. Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. International conference on machine learning. 2019;90:6105–14.
31. Liu R, Wang X, Wu Q, Dai L, Fang X, Yan T, et al. DeepDRiD: diabetic retinopathy-grading and image quality estimation challenge. Patterns (N Y). 2022;3(6):100512. https://doi.org/10.1016/j.patter.2022.100512 PMID: 35755875
32. Boreiko V, Ilanchezian I, Ayhan M, Muller S, Koch L, Faber H. Visual explanations for the detection of diabetic retinopathy from retinal fundus images. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2022. p. 539–49.
33. Mueller S, Heidrich H, Koch LM, Berens P. Fundus circle cropping. Available from: https://github.com/berenslab/fundus_circle_cropping
34. Huang Y, Lin L, Cheng P, Lyu J, Tang X. Identifying the key components in ResNet-50 for diabetic retinopathy grading from fundus images: a systematic investigation. arXiv preprint 2021. https://arxiv.org/abs/2110.14160

35. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR); 2016.
36. Ayhan MS, Kümmerle LB, Kühlewein L, Inhoffen W, Aliyeva G, Ziemssen F, et al. Clinical validation of saliency maps for understanding deep neural networks in ophthalmology. Med Image Anal. 2022;77:102364. https://doi.org/10.1016/j.media.2022.102364 PMID: 35101727
37. Ferrer L, Riera P. Confidence Intervals for evaluation in machine learning. Available from: https://github.com/luferrer/ConfidenceIntervals
38. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: Proceedings of the International Conference on Machine Learning. PMLR; 2017. p. 3319–28.
39. Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human-computer collaboration for skin cancer recognition. Nat Med. 2020;26(8):1229–34. https://doi.org/10.1038/s41591-020-0942-0 PMID: 32572267
40. Ng AY, Oberije CJG, Ambrózay É, Szabó E, Serfőző O, Karpati E, et al. Prospective implementation of AI-assisted screen reading to improve early detection of breast cancer. Nat Med. 2023;29(12):3044–9. https://doi.org/10.1038/s41591-023-02625-9 PMID: 37973948
41. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005 PMID: 28778026
42. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–2020): a comparative analysis. Lancet Digit Health. 2021;3(3):e195–203. https://doi.org/10.1016/S2589-7500(20)30292-2 PMID: 33478929
43. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. 2019;1(6):e271–97. https://doi.org/10.1016/S2589-7500(19)30123-2 PMID: 33323251
44. González-Gonzalo C, Thee EF, Klaver CCW, Lee AY, Schlingemann RO, Tufail A, et al. Trustworthy AI: Closing the gap between development and integration of AI systems in ophthalmic practice. Prog Retin Eye Res. 2022;90:101034. https://doi.org/10.1016/j.preteyeres.2021.101034 PMID: 34902546
45. Frasca M, La Torre D, Pravettoni G, Cutica I. Explainable and interpretable artificial intelligence in medicine: a systematic bibliometric review. Discov Artif Intell. 2024;4(1):15. https://doi.org/10.1007/s44163-024-00114-7
46. Chen C, Li O, Tao D, Barnett A, Rudin C, Su JK. This looks like that: deep learning for interpretable image recognition. Adv Neural Inf Process Syst. 2019:32.
47. Djoumessi K, Bah B, Kühlewein L, Berens P, Koch L. This actually looks like that: Proto-BagNets for local and global interpretability-by-design. Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2024. p. 718–28.
48. Dai L, Sheng B, Chen T, Wu Q, Liu R, Cai C, et al. A deep learning system for predicting time to progression of diabetic retinopathy. Nat Med. 2024;30(2):584–94. https://doi.org/10.1038/s41591-023-02702-z PMID: 38177850
49. Gervelmeyer J, Mueller S, Djoumessi K, Merle D, Clark S, Koch L. Interpretable-by-design deep survival analysis for disease progression modeling. In:International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2024. p. 502–12.