



OPEN

Interpretable detection of epiretinal membrane from optical coherence tomography with deep neural networks

Murat Seçkin Ayhan^{1,5}, Jonas Neubauer^{2,5}, Mehmet Murat Uzel^{2,3}, Faik Gelisken^{2✉} & Philipp Berens^{1,4✉}

This study aimed to automatically detect epiretinal membranes (ERM) in various OCT-scans of the central and paracentral macula region and classify them by size using deep-neural-networks (DNNs). To this end, 11,061 OCT-images were included and graded according to the presence of an ERM and its size (small 100–1000 μm , large > 1000 μm). The data set was divided into training, validation and test sets (75%, 10%, 15% of the data, respectively). An ensemble of DNNs was trained and saliency maps were generated using Guided-Backprob. OCT-scans were also transformed into a one-dimensional-value using t-SNE analysis. The DNNs' receiver-operating-characteristics on the test set showed a high performance for no-ERM, small-ERM and large-ERM cases (AUC: 0.99, 0.92, 0.99, respectively; 3-way accuracy: 89%), with small-ERMs being the most difficult ones to detect. t-SNE analysis sorted cases by size and, in particular, revealed increased classification uncertainty at the transitions between groups. Saliency maps reliably highlighted ERM, regardless of the presence of other OCT features (i.e. retinal-thickening, intraretinal pseudo-cysts, epiretinal-proliferation) and entities such as ERM-retinoschisis, macular-pseudohole and lamellar-macular-hole. This study showed therefore that DNNs can reliably detect and grade ERMs according to their size not only in the fovea but also in the paracentral region. This is also achieved in cases of hard-to-detect, small-ERMs. In addition, the generated saliency maps can be used to highlight small-ERMs that might otherwise be missed. The proposed model could be used for screening-programs or decision-support-systems in the future.

Epiretinal Membrane (ERM) is a common retinal disease, occurring mainly in elderly patients, with an incidence reported between 2.2 and 28.9%^{1–3}. It is characterized as an avascular fibro-cellular membrane on the innermost retinal layer and can lead to tractional changes and disruption of the retinal structure. Patients are often asymptomatic in early stages but reduced visual acuity, visual disturbances and increasing metamorphopsia are frequently seen in the later stages as the disease progresses⁴. So far, the only treatment is a vitrectomy with epiretinal peeling which increases visual acuity in most cases^{5,6}.

ERMs can be identified by funduscopy or on retinal fundus images; however, the gold standard nowadays for the diagnosis is Optical Coherence Tomography (OCT), where ERMs appear as a hyperreflective membrane on the inner surface of the retina. Here, also early stages of the disease and changes of the ERM over time are visible⁷. Owing to recent advances in deep learning and the large-scale analysis of medical images via deep neural networks (DNNs)^{8–10}, fundus and OCT imaging modalities have been investigated for automated ERM detection from retinal images^{11–15}. Nevertheless, these studies were limited to relatively small datasets or patients with ERM constituted only a small fraction of larger datasets^{11–15}. In addition, these studies primarily used data from patients with advanced stages of the disease, which are comparably easy to classify. For medical reasons,

¹Institute for Ophthalmic Research, University of Tübingen, Elfriede Aulhorn Str. 7, 72076 Tübingen, Germany. ²University Eye Clinic, University of Tübingen, Tübingen, Germany. ³Department of Ophthalmology, Balikesir University School of Medicine, Balikesir, Turkey. ⁴Tübingen AI Center, Tübingen, Germany. ⁵These authors contributed equally: Murat Seçkin Ayhan and Jonas Neubauer. ✉email: Faik.Gelisken@med.uni-tuebingen.de; philipp.berens@uni-tuebingen.de

it is similarly important to detect early and low-grade stages of ERMs. It is also of medical interest to segment ERMs in retinal images and some studies have automatically achieved this via DNNs trained on small data sets of 20 patients^{16–18}.

In this study, we investigated the automatic detection and classification of ERMs by developing an ensemble of DNNs with 11,061 clinically graded OCT-based images from 624 eyes (461 patients) and generated ensemble-based saliency maps to gain further insights into the decision-making process of the DNNs. Our model reliably detected small and large ERMs on both foveal and extrafoveal OCT scans with well calibrated uncertainty estimates, which is unique to our study.

Methods

Dataset

Our dataset consisted of 624 OCT volume scans from 624 eyes of 461 patients presenting to the Department of Ophthalmology at the University of Tübingen, resulting in a total of 11,061 images. The vast majority of patients were Caucasian. All OCT images were collected with Heidelberg Spectralis OCT (Heidelberg Engineering, Heidelberg, Germany) (Table 1). The majority of the scans were standardized, horizontal fovea centered volume scans containing 25 cross-sections with a distance between the B-Scans of 61 μm and a resolution of 384 \times 496 pixels. Approximately 3% of the images were single scans through the fovea (oblique, vertical or horizontal) with a larger width and were cropped for further analysis to a standardized width of 384 pixels, focused on the fovea. An ERM was defined as a hyperreflective membrane on the inner surface of the retina. Eyes with secondary ERM, high myopia (< -6 D), accompanying retinal diseases, such as diabetic maculopathy, retinal vein occlusion, advanced age-related macular degeneration, previous ocular trauma or vitrectomy, and poor-quality of OCT images were excluded.

All images were graded by a retina specialist according to the presence of an ERM and its size, dividing them into scans without ERM, with small ERM (100–1000 μm) and large ERM (> 1000 μm) (Fig. 1). The size of the ERM on each OCT scan was measured using a digital measurement tool and each image was classified individually. Therefore, one large membrane covering a larger area of the retina could be classified in OCT scans as small or large depending on the orientation and position of the scan. The dataset included also patients with features or entities associated with ERM, such as retinal thickening, intraretinal pseudocysts, epiretinal proliferation, ERM-retinoschisis, macular pseudohole and lamellar macular hole. To verify the dataset grading, we randomly sampled 500 B-scans representative of the three ERM classes as well as entities associated with ERM and had them re-graded by another retina specialist without disclosing the former specialist's ERM grades. Out of 500 B-scans, the specialists disagreed on only 19 images, which led to Cohen's kappa scores of 0.948 and 0.963 with linear and quadratic weighting schemes, respectively. A closer look into the grades of 19 B-scans revealed that the rare disagreements among graders were mostly (17 out of 19) between adjacent classes, i.e., No ERM vs Small ERM or Small ERM vs Large ERM. The graders diverged only on two B-scans by assigning No ERM or Large ERM labels in opposite scenarios. The classification of the OCT scans were performed by two ophthalmologists who had 11 and 9 years of experience in medical retina. The study was conducted according to the guidelines and standards of the Declaration of Helsinki, and was approved by the Ethics Committee of the University of Tübingen, Germany.

Network architecture and model development

We used the well-known ResNet50¹⁹ and InceptionV3²⁰ architectures implemented in Keras²¹ and pretrained on ImageNet²². We modified and fine-tuned them to our ERM classification tasks (Fig. 2). For each, we used max pooling and average pooling together at the end of the convolutional stack and concatenated their outputs. This combination had led to performance improvements in previous work^{23–26}. Additionally, we afterwards used two dense layers with 2048 and 512 units, which were followed by Batch Normalization²⁷ and ReLU activation²⁸. All weight layers except the penultimate layer were equipped with L_2 regularization. We employed L_1 regularization to promote sparsity in the penultimate layer. Finally, we replaced the classification layer with a binary classifier with sigmoid function for basic ERM detection or a 3-way softmax classifier for detection as well as classification of ERMs according to their size.

To train and evaluate our DNNs, we used random patient-based partitions of our data (Table 1). We trained them with cross-entropy loss on the training set: $\mathcal{D} = \{x_n, y_n\}_{n=1}^N$, where y_n is an

	Partition		
	Training (75%)	Validation (10%)	Test (15%)
Number of patients	345	47	69
Avg. age (95% conf. int.)	69.25 (68.35–70.14)	67.10 (64.46–69.74)	69.48 (67.29–71.67)
Number of images	8341	1146	1574
Laterality (L/R)	4174/4167	450/696	677/897
No ERM	3074 (36.9%)	436 (38.0%)	535 (34.0%)
Small ERM	1398 (16.8%)	204 (17.8%)	269 (17.1%)
Large ERM	3869 (46.3%)	506 (44.2%)	770 (48.9%)

Table 1. Summary of the data set. ERM, Epiretinal membrane.

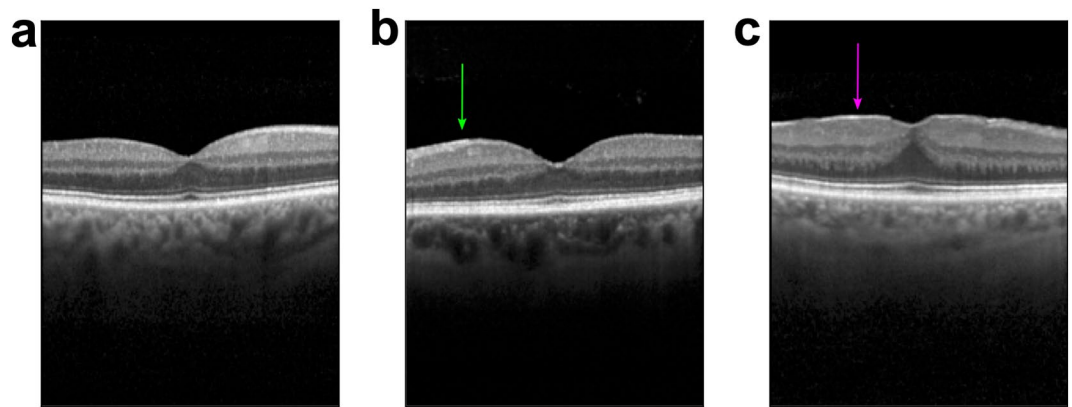


Figure 1. Exemplary optical coherence tomography images of the fovea. (a) No epiretinal membrane (ERM). (b) Small ERM (100–1000 μm) (green arrow). (c) Large ERM (> 1000 μm) (magenta arrow).

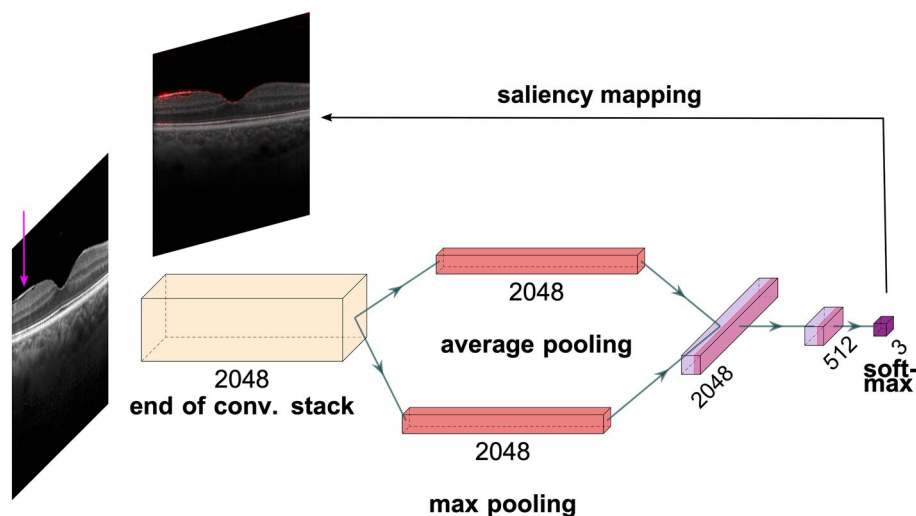


Figure 2. Schematic explanation of the ERM size analysis (plotted with PlotNeuralNet³¹): Given a retinal image with large ERM (indicated by an arrow in magenta), convolutional stack of the InceptionV3 architecture extracts 2048 feature maps. These are pooled with respect to average and max feature activations and fed into the fully connected (dense) layers with 2048 and 512 units. Finally, a 3-way softmax function achieves classification based on 512 features from the penultimate layer and detects large ERM. A saliency map highlights the regions relevant to the DNN decision.

expert-assigned label in binomial or multinomial (1-hot) representation for an image \mathbf{x}_n . Using the multinomial representation for the sake of generality, the average cross-entropy on the training data can be expressed as follows: $\mathcal{L}(\mathcal{D}, f_{\theta}(\cdot)) = \frac{1}{N} \sum_{n=1}^N l(y_n, f_{\theta}(x_n))$, where $f_{\theta}(\cdot)$ represents a DNN parameterized by θ , $l(y_n, f_{\theta}(x_n)) = -\sum_{k=1}^K y_{n,k} \log p_{n,k}$ and $p_{n,k}$ is a predicted class probability estimated via the softmax function for the k -th class out of $K=3$. For $K=2$, $l(y_n, f_{\theta}(x_n))$ is reduced to binary cross-entropy.

We countered the class imbalance in the data with random oversampling (Table 1). Using Stochastic Gradient Descent (SGD) with Nesterov's Accelerated Gradients (NAG)^{29,30}, the minibatch size of 16, a momentum coefficient of 0.9, an initial learning rate of 0.0001, a decay rate of 0.000001 and a regularization constant of 0.00001, we trained DNNs for at least 120 epochs (see in the next subsection for longer training).

During the first three epochs, convolutional stacks were frozen and only dense layers were trained. Then, all layers were fine-tuned to tasks. Models with the best validation accuracy were used on the test set.

Data augmentation and image preprocessing

To improve generalization to unseen data, we used *mixup*³² as a data augmentation technique. Given two examples (\mathbf{x}_i, y_i) and (\mathbf{x}_j, y_j) , *mixup* creates synthetic examples via their convex combinations:

$$\hat{\mathbf{x}} = \lambda \mathbf{x}_i + (1 - \lambda) \mathbf{x}_j, \quad \hat{y} = \lambda y_i + (1 - \lambda) y_j, \quad \lambda \in [0, 1].$$

Examples were randomly drawn from training data and $\lambda \sim \text{Beta}(\alpha, \alpha)$ for $\alpha \in (0, \infty)$. As $\alpha \rightarrow 0$, the effect of *mixup* diminishes. We used $\alpha \in [0, 0.1, 0.2, 0.3, 0.4]$ for 120, 120, 120, 150 and 200 epochs, respectively. In the warm-up period, we set $\alpha = 0$ for the first five epochs. In addition, we used standard data augmentation operations. The augmentation pipeline included random rotation within ± 45 degrees, horizontal and vertical translations within ± 30 pixels, brightness adjustments within $\pm 10\%$, zoom within $\pm 10\%$, and horizontal and vertical flips.

Overconfidence and calibration of predictive probabilities

DNNs trained with hard-coded labels and cross-entropy loss are often overconfident about their predictions^{33–36}, i.e. their predictive probabilities do not reliably indicate their expected accuracy. Label smoothing via *mixup* can already alleviate this miscalibration of DNNs³⁵. In addition to *mixup*, we used Deep Ensembles, which have been shown to improve the accuracy and calibration of DNNs^{37–39}. We constructed our ensembles with five DNNs trained for ERM classification (Supplementary Table 2), using the network architecture, hyperparameters and training procedures described above. DNNs were diversified by random initializations of dense layers, shuffling of training examples as well as mixing and data augmentation.

Embedding of images

To obtain insights into the feature representations learned by our ERM classification networks and their ensembles, we used t-Stochastic Neighborhood Embeddings (t-SNE)⁴⁰. t-SNE is a non-linear dimensionality reduction method, which also allows for the interpretation of high-dimensional data in low dimensions. To evaluate ensemble-based representations, we concatenated 512 features from ensemble members' penultimate layers and performed t-SNE based on 2560 features. We used *openTSNE*⁴¹ with PCA initialization to better preserve the global structure of the data and improve the reproducibility⁴². We ran the optimization for 1500 iterations with a perplexity of 200, Euclidean distance and an early exaggeration coefficient of 12 for the first 500 iterations.

Saliency maps

Saliency maps are a post-hoc interpretability technique, often used to generate explanations for DNN decisions^{43–47}. In the process, the prediction of a DNN is passed backwards through the network all the way back to the input image, where input pixels are associated with saliency scores according to their contribution to network outputs (Fig. 2). To compute saliency maps, we used the open-source library *iNNvestigate*⁴⁸ with the Guided Backprob algorithm⁴⁹. This algorithm has been evaluated for clinical relevance in ophthalmology and shown to perform consistently well across different network architectures, imaging modalities, such as retinal photography and OCT, and diagnostic scenarios involving diabetic retinopathy (DR), choroidal neovascularization (CNV), diabetic macular edema (DME) or neovascular age-related macular degeneration (nAMD)^{26,50,51}.

Results

We trained DNNs to detect and classify ERMs from 11,061 individual B-scans extracted from 624 OCT volume scans. The B-scans included images with no ERMs, small ERMs and large ERMs (Fig. 1). For the binary ERM detection task, we grouped small and large ERMs together. During training, we used a recent data augmentation and regularization technique called *mixup*, along with standard data augmentation operations. The ERM detection performance of DNNs were maintained or marginally improved with the degree of mixing and longer training (Supplementary Table 1). In addition to *mixup*, we constructed ensembles of DNNs but this also maintained the ERM detection performance across both DNN architectures, ResNet50 and InceptionV3. Given the comparable ERM detection performances of these two well-known architectures and widespread adoption of the latter for medical imaging^{26,52–54}, we used InceptionV3 for both detection and classification of ERMs according to their size. Interestingly, the performances of DNNs improved with the degree of mixing and longer training in this more challenging and clinically relevant scenario (Supplementary Table 2). To further improve performance, we used ensembles of DNNs once again (Supplementary Table 2).

Our 3-way classification DNNs accurately detected the presence of ERM in retinal images and were also able to determine the ERM size, which is an important feature of clinical relevance, with high accuracy (Supplementary Table 2, Fig. 3). The best ensemble model was obtained from DNNs trained with intense mixing ($\alpha = 0.4$) for 200 epochs, achieving a 3-way classification accuracy of 89.33% (indicated by the gray row in Supplementary Table 2) and an AUC of 0.99, 0.92 and 0.99 for normal B-scans, B-scans with small and large ERMs, respectively. Interestingly, small ERMs were more difficult for DNNs to detect than the large ones (Fig. 3a,b). While this difficulty can be attributed to the ERM pathophysiology that causes small ERMs to be the most difficult to identify among the three classes, ablation of ensembling along with *mixup* led to inferior performance and impacted the small ERM detection rate most (see supplementary Fig. 1a,b).

We also assessed the uncertainty estimates provided by our ensemble model and found that its predictive probabilities for the training and validation data were slightly oversmoothed, likely due to the combined effects of *mixup* and ensembling, but overall well-calibrated on the test data with a small adaptive expected calibration error of 0.02 (Fig. 3c). With the ablation of ensembling and *mixup*, the calibration error was 0.05 (supplementary Fig. 1c). This indicated that the uncertainty estimates reported by our ensemble with *mixup* can be expected to provide useful information about the performance of the model, with high uncertainties corresponding to more errors, an information which may be taken into account by clinicians making decisions based on the DNNs outputs.

To better understand which areas of the images were important to the DNN's decision-making process, we created saliency maps using the Guided Backprob algorithm⁴⁹. These saliency maps clearly highlighted important areas of interest mainly on the inner surface of the retina (Fig. 4d–l), including scans with a small ERM. This is

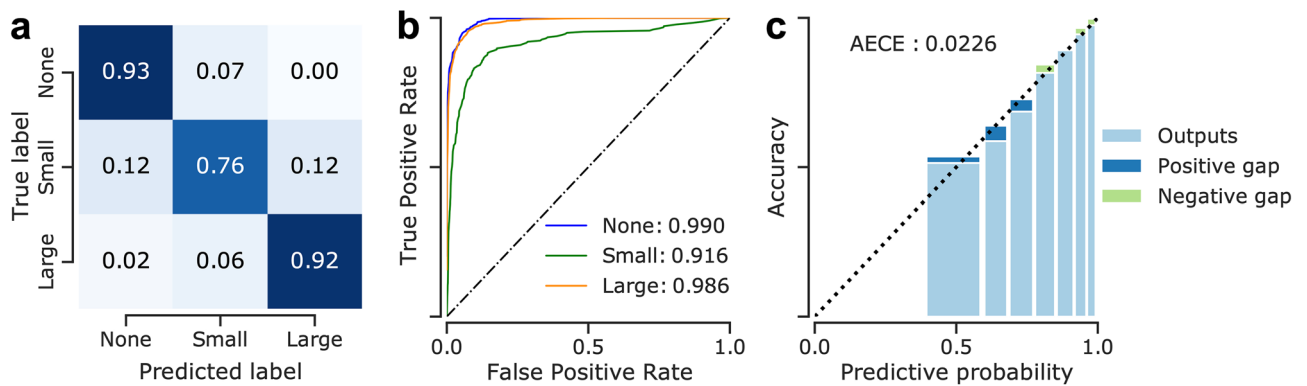


Figure 3. Detailed analysis of classification performance on the test set. **(a):** Confusion matrix for the selected ensemble model. **(b):** Receiver operating characteristics of the ensemble model. Numbers indicate AUC scores for the respective classes. **(c):** Reliability diagrams and calibration of the selected ensemble. The calibration error was estimated via reliability diagrams^{33,55,56} and the Adaptive Expected Calibration Error (AECE)⁵⁵. While a negative gap (light green) between predictive probability and accuracy indicates a lack of confidence in predictions, a positive gap (dark blue) points at overconfidence.

remarkable, since many of the small ERMs are very fine structures which are difficult to detect even for ophthalmologists. It should also be emphasized that associated retinal changes and entities (i.e. intraretinal pseudocysts, retinoschisis, lamellar macular hole) were not markedly highlighted on the saliency maps and thus seemed to be not crucial for the decision-making process of the DNN. In the cases of images without ERM, vitreous opacities were frequently marked (Fig. 4a–c).

The proposed model classified 169 images of the test set (1574 images) differently than the retina specialist (Fig. 5). Re-analysis of these images by two additional retina specialists revealed that 53 scans had indeed been correctly classified by the DNN and that most of these grader misclassifications were due to obvious documentation errors (for example, 24 of the 53 images originated from one volume scan of one patient with prominent ERM and were graded as "small ERM" by the human). In the clinical context, however, it is far more interesting to see how often the algorithm missed an ERM, which was the case in 62 scans. Upon inspection, we found that these were mostly very fine and small ERMs at the edge of what was considered as ERM present ($> 100 \mu\text{m}$) (Fig. 5a).

To study the representation learned by the DNNs, we embedded all images in our dataset into one dimension using t-Stochastic Neighborhood Embedding (t-SNE) based on the network activations (features) in the penultimate layers of the ensemble members (Fig. 6, also see Methods). In this representation, each dot corresponds to a single OCT B-scan, with the color indicating its class. We found that the resulting t-SNE map (Fig. 6a) followed the disease continuum with great accuracy ordering the discrete classes accordingly with ERM-negative cases being placed to the left and positive ones towards right. Pairing the t-SNE coordinates with predictive uncertainty associated with retinal images, we also found that the average uncertainty was highest at the boundaries between the stages of ERM (Fig. 6a). High uncertainty was also indicative of wrong predictions (Fig. 6b), in line with the finding that most misclassifications occurred at the transition between ERM-negative and small ERMs, and the good calibration of the model. In fact, incorrect and correct predictions were coupled with significantly higher and lower uncertainty, respectively ($p < 0.0001$, Mann-Whitney U test, $n_{\text{wrong}} = 446$, $n_{\text{correct}} = 10,615$) (Fig. 6c).

Discussion

We showed that DNNs can reliably detect ERMs of different severity on OCT images of the macula, classify them based on their size and provide well-calibrated uncertainty estimates for their decisions. In addition, to gain further insights into the DNNs' inner workings leading to ERM decisions, we computed ensemble-based saliency maps²⁶ that labeled ERMs independently of accompanying retinal changes with high confidence.

As the OCT exam provides a detailed visualization of the posterior pole and the retinal layers, various ERM classification systems have been proposed. Some distinguish between attached and detached vitreous, central retinal thickness or classify patients based on fovea involvement and changes in retinal layers^{4,57,58}. However, these grading systems are only applicable to OCT images of the fovea and have not gained general acceptance in the clinical routine. As the aim of our work was to provide an automatic system for robust detection of ERM, we therefore included not only single fovea scans but also OCT images from the para- and perifoveal region. In order to classify images with ERM further, we decided to group them by size as larger ERMs tend to alter retinal layers more severely than smaller ERMs. We developed an ensemble of DNNs to automatically detect and grade ERMs on foveal and extrafoveal OCT scans and our proposed model showed high accuracy in this multi-class classification scenario. In this respect, our study makes progress in follow-up to the existing work on ERM detection from retinal images, which so far considered binary scenarios, i.e., deciding either the absence or presence of ERM^{11–15}. In fact, our proposed DNN seems to grade the extent of the disease automatically. Despite being presented only categorical labels and having no explicit knowledge of the ERM pathology or development, the generated one dimensional t-SNE map (Fig. 6a) indicates that the model faithfully restored the disease continuum from data and ordered the classes accordingly with ERM-negative cases being placed to the left, small ERMs in

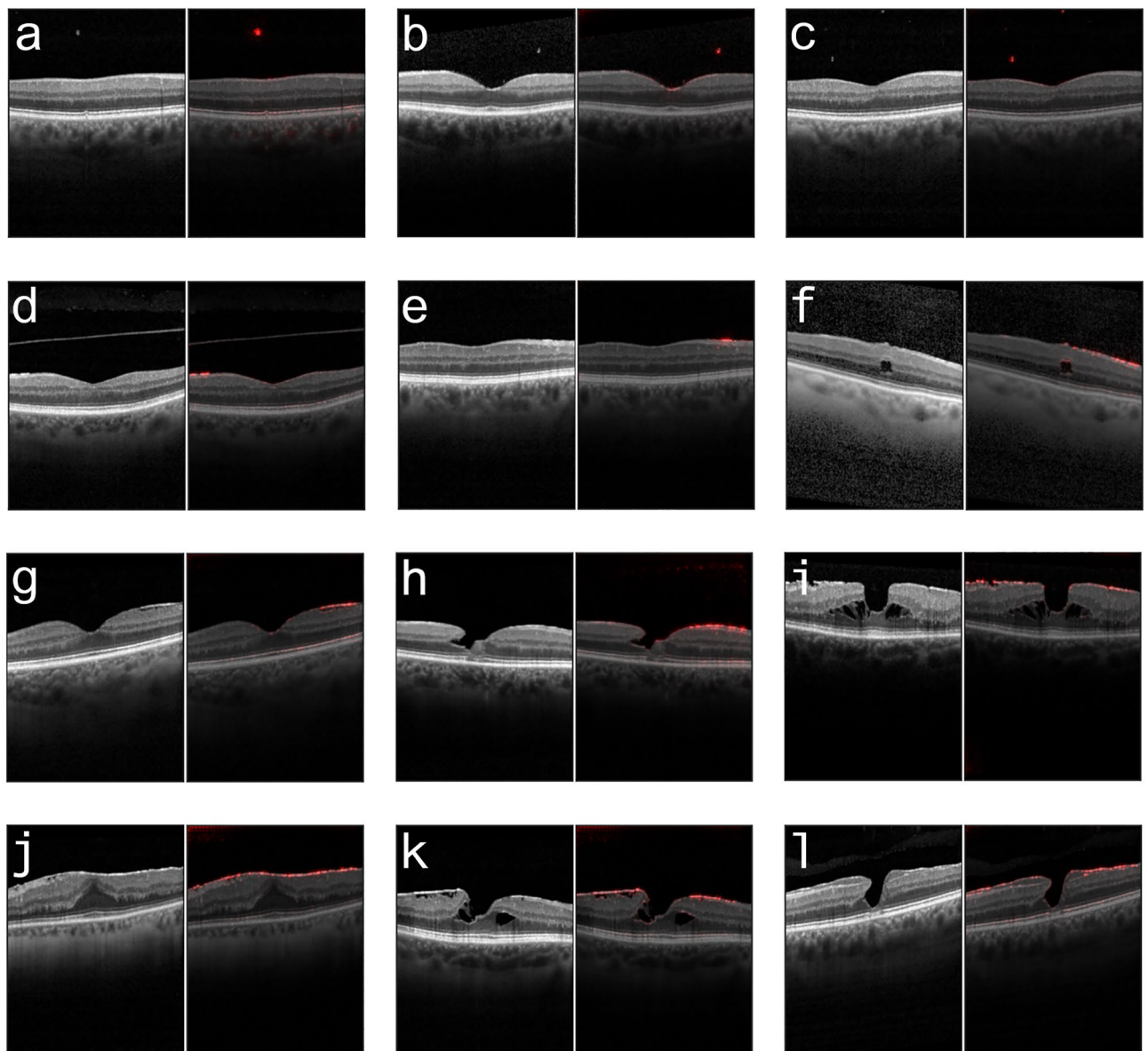


Figure 4. Examples of the correctly classified optical coherence tomography (OCT) images (left) and generated saliency maps (right), showing reliable highlighting of ERM in foveal and extra-foveal areas. **(a–c):** OCT scans without ERM and highlighted opacities in the vitreous. **(d):** Foveal scan with a small extrafoveal ERM. **(e):** Extrafoveal scan with a small ERM. **(f):** Extrafoveal scan with a large ERM and an intraretinal pseudocyst. **(g):** Foveal scan with a large extrafoveal ERM not affecting the foveal depression. **(h):** Lamellar macular hole with a large ERM. **(i):** A large ERM and foveoschisis. **(j):** Foveal scan with a large ERM and an elevated foveal depression. **(k):** Foveoschisis with a large ERM. **(l):** Early stage of a lamellar macular hole with a large ERM and epiretinal proliferation.

the center and large ERMs towards the right. Similar characteristics have been observed in the past for automatic ordering of diabetic retinopathy stages²⁵.

Our DNNs misclassified only a few large ERM scans as healthy, demonstrating that a more pronounced disease is easier to recognize on OCT scans. Furthermore, it should be noted that these misclassified scans presented only as a very fine hyperreflective line and could have been easily missed even by ophthalmologists. As one would therefore expect, small ERMs were more difficult for our DNNs to detect, compared to the large ones or cases of no ERM. Belonging to a transitional state between no ERM and advanced stages, the examples of small ERM also demonstrated an overlap with the other two classes. The predictive uncertainty of the proposed model was indicative of such difficult cases mostly located around decision boundaries. Interestingly, the peak uncertainty was observed when transitioning from ERM-negative to small ERM scans, indicating the difficulty of detecting such early-onset cases, similar to the mild DR cases reported previously²⁵. Distinguishing between adjacent stages seems to be not only for DNNs but also for retina specialists challenging, as the vast majority of the few disagreements among the two graders were between adjacent stages.

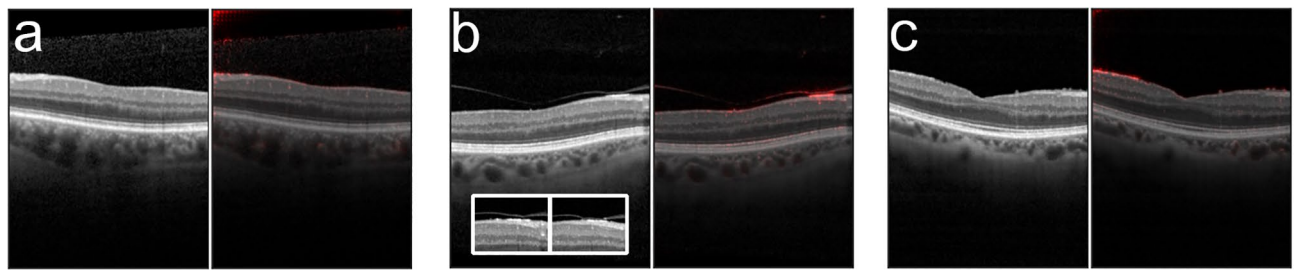


Figure 5. Examples of incorrectly classified optical coherence tomography (OCT) scans (left) and generated saliency maps (right). **(a):** This scan was classified as “no ERM”, while a small ERM of about 170 μm size is detectable. **(b):** Even though the ERM is highlighted in the saliency map the model classified this image incorrectly. The neighbouring OCT images superior and inferior (presented in the box) to this one were classified correctly. **(c):** A large ERM that has been classified by the model as a small ERM.

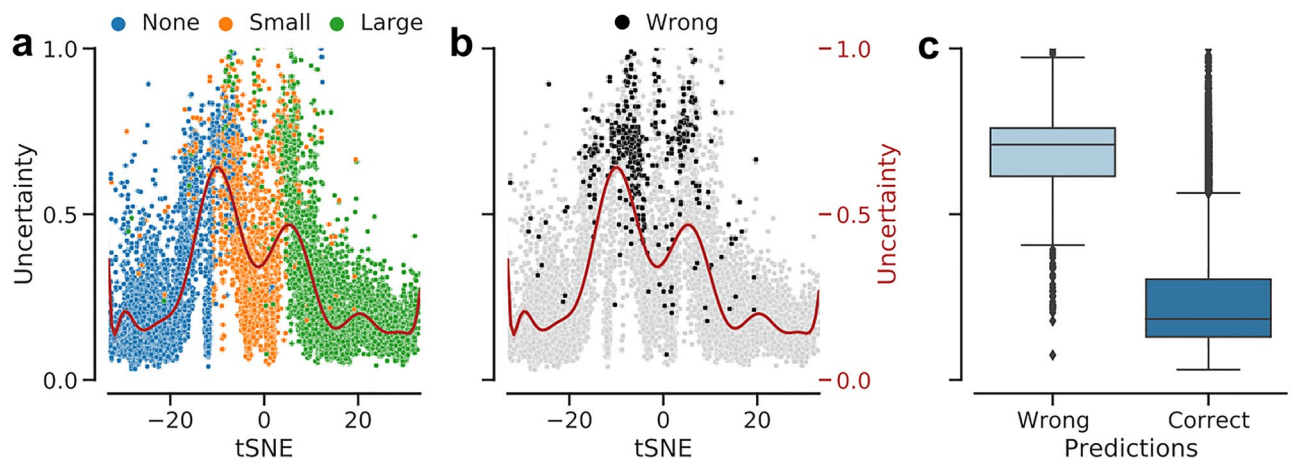


Figure 6. 1D t-SNE map via the best ensemble model and predictive uncertainty difference between the groups of wrong and correct predictions. The paired t-SNE coordinates with predictive uncertainty measured by the entropy of the ensemble output and fitted a polynomial regression curve to data. **(a):** Training, validation and test data aligned together and colored with respect to the ERM labels. **(b):** Same map as in (a) but colored w.r.t. correct (gray) and wrong (black) predictions. **(c):** Uncertainty distribution in groups of wrong and correct predictions.

The accuracy claims of other studies have to be interpreted therefore in light of the data used to train and evaluate their models, typically omitting small ERMs. Previous publications have often not provided much information on the data used, with regards to the severity of the ERM or OCT characteristics, but the shown examples indicate more advanced cases. Only Parra-Mora et al.¹⁵ stated that they used OCT scans of different disease severity, without further specifying the distribution of the stages. In contrast, the distribution of ERMs in the study presented here allows for a more accurate representation in terms of disease severity of the study population, as we collected a large dataset derived from the clinical routine of our clinic. Consequently, we covered the whole spectrum of different ERMs presenting with features (i.e. retinal thickening, intraretinal pseudocysts, retinoschisis) and other entities (i.e. ERM-retinoschisis, lamellar macular hole, macular pseudohole) associated with ERM. We implemented not only standardized horizontal OCT scans but also vertical and oblique oriented ones. In addition to treating ERM detection as a classification task as in this study, segmentation algorithms have also been trained to detect ERMs^{16–18}. Some of these works have primarily focused on the detection of the internal limiting membrane (ILM), which might be erroneous in eyes with a visible vitreous or posterior vitreous limiting membrane—as regularly seen in patients with ERM (i.e. Figs. 4d, 5b).

Our study presents a path for a robust detection of ERMs in a variety of patients, with different accompanying features and with different scan patterns. Also, explanations of our DNNs using saliency maps have highlighted the usefulness of the model in detecting ERMs, even if they are small and not located in the fovea. Vitreous opacities were frequently marked in images without ERM (i.e. Fig. 4a–c). This can be interpreted in accordance with the fact that saliency maps often appear diffuse and not well localized in the absence of pathologies which are supposed to be detected by the DNN. Previous ERM classification studies have relied on the GradCAM algorithm⁵⁹ to generate saliency maps which highlighted the approximate region of the retina where the ERM was detected^{12,15}. In contrast, our proposed saliency maps based on GuidedBackProp^{26,49} presents substantially more detail and a much finer resolution, while reliably detecting the ERM at the same time. Interestingly, saliency maps repeatedly marked outer retinal layers like the retinal pigment epithelium under the ERMs as important areas (Fig. 4). A possible explanation for this could be that the ERM casts a light shadow over the underlying retinal

structures, which could be detected by the model and used as an indirect OCT retinal biomarker. The precise representation of ERMs in the presented saliency maps could help to increase trust in the described DNN as it makes the decision-making process of the algorithm more understandable for patients and physicians. To this end, visual counterfactual explanations (VCEs) via adversarially robust ensembles⁶⁰ can be also used to augment saliency maps with more realistic and reliable visualizations in future studies.

Due to the increasing age of the general population it will be challenging to provide comprehensive ophthalmological care in the near future, thus shifting the focus to decision support systems or automatic screening approaches⁶¹. This is particularly the case for ERMs, since it usually starts as a monocular disease that is not always directly noticed by the patient and is probably associated with a worse visual prognosis in case of delayed treatment⁶². At the same time, the amount of data generated per patient and visit increases over the years, making it more difficult for a physician to accurately assess all of it. Thus, our state-of-the-art DNN could help ophthalmologists in the sense of a decision-support system and, for instance, point out abnormal areas of volume scans in order to prevent medical misdiagnosis and improve the quality of healthcare. Multi-task models are also promising for more comprehensive systems that will take into account multiple pathologies associated with ERM, as recently demonstrated for the case of nAMD activity detection⁵⁴.

A limitation of the current study, however, can be thought of as the position and orientation of OCT scans, which could have led to different ERM size measurements under different settings. To address this, an alternative would be to display the entire ERM on one OCT volume scan, segment it and finally measure the largest diameter or the area covered by the ERM. However, this is, due to technical reasons, in many cases not possible and it would be practically impossible to obtain these labels for such a large data set in a clinical setting. A possible avenue to explore in this regard is the use of self-supervised learning and foundation models⁶³ that would benefit from broad data and sidestep the need for expert labels.

Data availability

The datasets generated and analyzed during the current study are not publicly available due to privacy reasons but are available from the corresponding author on reasonable request.

Received: 12 December 2022; Accepted: 21 March 2024

Published online: 11 April 2024

References

- Ng, C. H. *et al.* Prevalence and risk factors for epiretinal membranes in a multi-ethnic united states population. *Ophthalmology* **118**, 694–699 (2011).
- Xiao, W. *et al.* Prevalence and risk factors of epiretinal membranes: A systematic review and meta-analysis of population-based studies. *BMJ Open* **7**, e014644 (2017).
- You, Q., Xu, L. & Jonas, J. Prevalence and associations of epiretinal membranes in adult Chinese: The Beijing eye study. *Eye* **22**, 874–879 (2008).
- Govetto, A., Lalane, R. A. III., Sarraf, D., Figueroa, M. S. & Hubschman, J. P. Insights into epiretinal membranes: Presence of ectopic inner foveal layers and a new optical coherence tomography staging scheme. *Am. J. Ophthalmol.* **175**, 99–113 (2017).
- Dawson, S., Shunmugam, M. & Williamson, T. Visual acuity outcomes following surgery for idiopathic epiretinal membrane: An analysis of data from 2001 to 2011. *Eye* **28**, 219–224 (2014).
- Massin, P. *et al.* Optical coherence tomography of idiopathic macular epiretinal membranes before and after surgery. *Am. J. Ophthalmol.* **130**, 732–739 (2000).
- Delyfer, M.-N. *et al.* Prevalence of epiretinal membranes in the ageing population using retinal colour images and SD-OCT: The Alienor Study. *Acta Ophthalmol.* **98**, e830–e838 (2020).
- Topol, E. J. High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.* **25**, 44 (2019).
- Esteve, A. *et al.* Deep learning-enabled medical computer vision. *NPJ. Digit. Med.* **4**, 1–9 (2021).
- Liu, X. *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* **1**, e271–e297 (2019).
- Lu, W. *et al.* Deep learning-based automated classification of multi-categorical abnormalities from optical coherence tomography images. *Transl. Vis. Sci. Technol.* **7**, 41–41 (2018).
- Lo, Y.-C. *et al.* Epiretinal membrane detection at the ophthalmologist level using deep learning of optical coherence tomography. *Sci. Rep.* **10**, 1–8 (2020).
- Son, J. *et al.* Development and validation of deep learning models for screening multiple abnormal findings in retinal fundus images. *Ophthalmology* **127**, 85–94 (2020).
- Casado-Garcia, Á. *et al.* Prediction of epiretinal membrane from retinal fundus images using deep learning. In *Conference of the Spanish Association for Artificial Intelligence* (2021), 3–13.
- Parra-Mora, E., Cazañas-Gordon, A., Proença, R. & da Silva Cruz, L. A. Epiretinal membrane detection in optical coherence tomography retinal images using deep learning. *IEEE Access* **9**, 99201–99219 (2021).
- Gende, M., De Moura, J., Novo, J., Charlón, P. & Ortega, M. Automatic segmentation and intuitive visualisation of the epiretinal membrane in 3D OCT images using deep convolutional approaches. *IEEE Access* **9**, 75993–76004 (2021).
- Gende, M., de Moura, J., Novo, J. & Ortega, M. End-to-end multi-task learning approaches for the joint epiretinal membrane segmentation and screening in OCT images. *Comput. Med. Imaging Graph.* **98**, 102068 (2022).
- Baamonde, S., de Moura, J., Novo, J., Charlón, P. & Ortega, M. Automatic identification and intuitive map representation of the epiretinal membrane presence in 3D OCT volumes. *Sensors* **19**, 5269 (2019).
- He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2016).
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), 2818–2826.
- Chollet, F. *et al.* Keras <https://github.com/fchollet/keras>.
- Russakovsky, O. *et al.* ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis. (IJCV)* **115**, 211–252 (2015).
- Leibig, C., Allken, V., Ayhan, M. S., Berens, P. & Wahl, S. Leveraging uncertainty information from deep neural networks for disease detection. *Sci. Rep.* **7**, 17816 (2017).
- Ayhan, M. S. & Berens, P. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *Proceedings of the International Conference on Medical Imaging with Deep Learning* (2018).

25. Ayhan, M. S. et al. Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection. *Medical Image Analysis*, 101724 (2020).
26. Ayhan, M. S. et al. Clinical validation of saliencymaps for understanding deep neural networks in ophthalmology. *Med. Image Anal.* **77**, 102364 (2022).
27. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning* (2015), 448–456.
28. Nair, V. & Hinton, G. E. *Rectified Linear Units Improve Restricted Boltzmann Machine* (Omnipress, Haifa, Israel, 2010), 807–814
29. Nesterov, Y. E. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk. SSSR* **269**, 543–547 (1983).
30. Sutskever, I., Martens, J., Dahl, G. E. & Hinton, G. E. On the importance of initialization and momentum in deep learning. In *ICML* (3) **28**, 5 (2013).
31. Iqbal, H. PlotNeuralNet Accessed: 2023 May 14. 2018. <https://github.com/HarisIqbal88/PlotNeuralNet>.
32. Zhang, H., Cisse, M., Dauphin, Y. N. & Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations* (2018).
33. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (2017), 1321–1330.
34. Vaicenavicius, J. et al. Evaluating model calibration in classification. In *Proceedings of Machine Learning Research* (eds Chaudhuri, K. & Sugiyama, M.) **89** (PMLR, Apr. 2019), 3459–3467.
35. Thulasidasan, S., Chennupati, G., Bilmes, J. A., Bhattacharya, T. & Michalak, S. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Adv. Neural Inf. Process. Syst.* **32** (2019).
36. Meinke, A. & Hein, M. Towards neural networks that provably know when they don't know in *8th International Conference on Learning Representations, ICLR 2020* (OpenReview.net, 2020). <https://openreview.net/forum?id=ByxGkySKwH>.
37. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems* (2017), 6405–6416.
38. Fort, S., Hu, H. & Lakshminarayanan, B. Deep ensembles: A loss landscape perspective. <http://arxiv.org/abs/arXiv:1912.02757> (2019).
39. Ovadia, Y. et al. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems* (2019), 13991–14002.
40. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
41. Poliar, P. G., Stražar, M. & Zupan, B. openTSNE: A modular python library for t-SNE dimensionality reduction and embedding. *bioRxiv* **3**, 861 (2019).
42. Kobak, D. & Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.* **10**, 1–14 (2019).
43. Quellec, G., Charrière, K., Boudi, Y., Cochener, B. & Lamard, M. Deep image mining for diabetic retinopathy screening. *Med. Image Anal.* **39**, 178–193 (2017).
44. Ancona, M., Ceolini, E., Öztireli, C. & Gross, M. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings* (OpenReview.net, 2018). <https://openreview.net/forum?id=Sy21R9JAW>.
45. Montavon, G., Samek, W. & Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.* **73**, 1–15 (2018).
46. Montavon, G., Binder, A., Lapuschkin, S., Samek, W. & Müller, K.-R. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning 193–209* (Springer, Cham, 2019).
47. Reyes, M. et al. On the interpretability of artificial intelligence in radiology: Challenges and opportunities. *Radiol. Artif. Intell.* **2**, 190043 (2020).
48. Alber, M. et al. iNNvestigate neural networks. *J. Mach. Learn. Res.* **20**, 1–8 (2019).
49. Springenberg, J. T., Dosovitskiy, A., Brox, T. & Riedmiller, M. Striving for simplicity: The all convolutional net. <http://arxiv.org/abs/arXiv:1412.6806> (2014).
50. Singh, A. et al. What is the optimal attribution method for explainable ophthalmic disease classification? In *Ophthalmic Medical Image Analysis* (eds Fu, H. et al.) 21–31 (Springer, Cham, 2020).
51. Van Craenendonck, T., Elen, B., Gerrits, N. & De Boever, P. Systematic comparison of heatmapping techniques in deep learning in the context of diabetic retinopathy lesion detection. *Transl. Vis. Sci. Technol.* **9**, 64–64 (2020).
52. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115 (2017).
53. Poplin, R. et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* **2**, 158–164 (2018).
54. Ayhan, M. S. et al. Multitask learning for activity detection in neovascular age-related macular degeneration. *Transl. Vis. Sci. Technol.* **12**, 12–12 (2023).
55. Ding, Y., Liu, J., Xiong, J. & Shi, Y. Evaluation of neural network uncertainty estimation with application to resource-constrained platforms. <http://arxiv.org/abs/arXiv:1903.02050> (2019).
56. Niculescu-Mizil, A. & Caruana, R. Predicting good probabilities with supervised learning. In *Proceedings of the 22Nd International Conference on Machine Learning* (ACM, Bonn, Germany, 2005), 625–632.
57. Konidaris, V., Androudi, S., Alexandridis, A., Dastiridou, A. & Brazitikos, P. Optical coherence tomography guided classification of epiretinal membranes. *Int. Ophthalmol.* **35**, 495–501 (2015).
58. Hwang, J.-U. et al. Assessment of macular function for idiopathic epiretinal membranes classified by spectral-domain optical coherence tomography. *Investig. Ophthalmol. Vis. Sci.* **53**, 3562–3569 (2012).
59. Selvaraju, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), 618–626.
60. Boreiko, V. et al. Visual explanations for the detection of diabetic retinopathy from retinal fundus images. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part II* (2022), 539–549.
61. Resnikoff, S. et al. Estimated number of ophthalmologists worldwide (International Council of Ophthalmology update): Will we meet the needs?. *Br. J. Ophthalmol.* **104**, 588–592 (2020).
62. Rahman, R. & Stephenson, J. Early surgery for epiretinal membrane preserves more vision for patients. *Eye* **28**, 410–414 (2014).
63. Zhou, Y. et al. A foundation model for generalizable disease detection from retinal images. *Nature* **622**(7981), 156–163 (2023).

Acknowledgements

We acknowledge support from the Open Access Publication Fund of the University of Tübingen. We thank the German Ministry of Science and Education (BMBF) for funding through the Tübingen AI Center (FKZ 01IS18039A) and the German Science Foundation for funding through a Heisenberg Professorship (BE5601/4-2) and the Excellence Cluster "Machine Learning—New Perspectives for Science" (EXC 2064, project number 390727645). MMU received a postdoctorate international research grant (grant number: BIDEB-2219) from

“The Scientific and Technological Research Council of Turkey—TUBITAK.” The funding bodies did not have any influence in the study planning and design.

Author contributions

All authors planned the study. M.S.A. and J.N. wrote the main manuscript text. M.S.A. and P.B. designed the algorithm. M.M.U., J.N. and F.G. curated data. All authors interpreted and analyzed the data and reviewed the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-57798-1>.

Correspondence and requests for materials should be addressed to F.G. or P.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024