PNAS Nexus, 2023, **2**, 1–14
https://doi.org/10.1093/pnasnexus/pgad290
Advance access publication 6 September 2023
Research Report

Artificial intelligence, explainability, and the scientific method: A proof-of-concept study on novel retinal biomarker discovery

Parsa Delavari[a,b,1], Gulcenur Ozturan[a,1], Lei Yuan[a], Özgür Yilmaz[c] and Ipek Oruc[a,b,*]

[a]Ophthalmology and Visual Sciences, University of British Columbia, Vancouver, V5Z 0A6 BC, Canada
[b]Neuroscience, University of British Columbia, Djavad Mowafaghian Centre for Brain Health, Vancouver, V6T 1Z3 BC, Canada
[c]Mathematics, University of British Columbia, Vancouver, V6T 1Z2 BC, Canada
*To whom correspondence should be addressed: Email: ipor@mail.ubc.ca
[1]P.D. and G.O. contributed equally to this work.
Edited By: Rui Reis

Abstract

We present a structured approach to combine explainability of artificial intelligence (AI) with the scientific method for scientific discovery. We demonstrate the utility of this approach in a proof-of-concept study where we uncover biomarkers from a convolutional neural network (CNN) model trained to classify patient sex in retinal images. This is a trait that is not currently recognized by diagnosticians in retinal images, yet, one successfully classified by CNNs. Our methodology consists of four phases: In Phase 1, *CNN development*, we train a visual geometry group (VGG) model to recognize patient sex in retinal images. In Phase 2, *Inspiration*, we review visualizations obtained from post hoc interpretability tools to make observations, and articulate exploratory hypotheses. Here, we listed 14 hypotheses retinal sex differences. In Phase 3, *Exploration*, we test all exploratory hypotheses on an independent dataset. Out of 14 exploratory hypotheses, nine revealed significant differences. In Phase 4, *Verification*, we re-tested the nine flagged hypotheses on a new dataset. Five were verified, revealing (i) significantly greater length, (ii) more nodes, and (iii) more branches of retinal vasculature, (iv) greater retinal area covered by the vessels in the superior temporal quadrant, and (v) darker peripapillary region in male eyes. Finally, we trained a group of ophthalmologists ($N = 26$) to recognize the novel retinal features for sex classification. While their pretraining performance was not different from chance level or the performance of a nonexpert group ($N = 31$), after training, their performance increased significantly ($p < 0.001$, $d = 2.63$). These findings showcase the potential for retinal biomarker discovery through CNN applications, with the added utility of empowering medical practitioners with new diagnostic capabilities to enhance their clinical toolkit.

Keywords: artificial intelligence, convolutional neural networks, retinal fundus images, retinal biomarkers, medical image perception

Significance Statement

AI is often utilized to automate tasks typically carried out by human experts. Can we also harness the power of AI to enhance their expertise? Current challenges in explainability of AI remains a barrier. Here, we introduce a methodology that utilizes the scientific method to augment existing explainability tools. We showcase the potential of this method on a medical image recognition application, where we discover novel retinal markers. We also show that it is possible to teach clinicians to identify subtle, yet critical, novel features, ultimately enhancing their ability to provide accurate diagnoses. The proposed methodology presents a promising avenue for identifying novel retinal biomarkers associated with a variety of conditions, including neurodegenerative and cardiovascular diseases.

Introduction

Deep neural networks (DNNs) are powerful machine learning models performing tasks previously considered to be exclusive to humans (1), such as visual object and face recognition, image segmentation, and natural language processing. Convolutional neural networks (CNNs) are a subtype of deep learning networks specialized for image processing and classification. Relying on their superior performance, these models have found their way to almost every field of study, and medical imaging is no exception. CNNs are applied to various imaging modalities, such as computed tomography, magnetic resonance imaging, X-ray, electrocardiogram, and dermoscopic images to facilitate

Competing Interest: The authors declare no competing interest.
Received: June 20, 2023. **Accepted:** August 28, 2023

computer-aided detection of abnormalities and to assist clinicians in the process of disease diagnosis and management (2–5).

Deep learning models have been widely used in retinal imaging modalities (6, 7). Recently, CNNs have seen tremendous success in predicting eye diseases such as diabetic retinopathy (DR) (8), age-related macular degeneration (AMD) (9, 10), and glaucoma (11) based on retinal fundus photographs. In addition to various signs of ocular diseases, features, and traits beyond eye health are also visible to artificial intelligence (AI) in this imaging modality. For example, Poplin et al. (12) successfully predicted cardiovascular risk factors, including patient age, smoking status, hemoglobin A1c, systolic blood pressure, diastolic blood pressure, and body mass index. Other systemic biomarkers, such as creatinine level and body composition indices (muscle mass, height, weight) can also be predicted by CNNs based on fundoscopic images (13). Furthermore, the retina is considered a window to the brain, being the only part of the central nervous system that can be observed directly and noninvasively. The surface of the retina is covered by retinal ganglion cells, a subtype of neurons, and it shares many anatomical, physiological, and thus, pathological properties with the brain. Several studies have shown the effects of Alzheimer's disease (AD) on the retina, bringing forward the potential for the development of new early diagnosis methodologies (14–18).

Despite the successful application of AI in a wide range of medical image classification tasks, the lack of transparency in its output decisions has been a roadblock to its widespread adoption in medicine (19). Encapsulating the relationship between the input image and the output label is challenging in these so-called black-box models. CNNs, in particular, consist of numerous layers and tens of millions of parameters. The inherent complexity in this family of models has led to the emergence of a separate field of study, "deep learning explainability," in search for human-comprehensible interpretability tools to shed light on the underlying decision-making processes of the networks. Saliency maps, also known as heat maps or attention maps, are one of the most common interpretability tools for CNNs, and are used to highlight the image areas that have the greatest contribution to the model's decision (20–22). These saliency maps, however, offer only superficial information limited to the spatial distribution of regions used by the network, without a precise description of what features within the highlighted regions contributed to the model's decision.

The other popular family of CNN explainability tools is feature visualization, which aims to uncover the visual features learned by the network, i.e. the preferred stimuli of a particular neuron, layer, or final label in the network (22–24). However, because of the complex nature of CNNs, a wide range of regularization techniques are needed to obtain meaningful and human-interpretable visualizations (23). Furthermore, the insights gained from feature visualizations are inherently dependent on subjective impressions, rendering them susceptible to variability and imprecision [e.g. (25, 26)]. This problem is relevant for any classification tasks where human observers perform comparably with the models, but especially so for tasks in which human experts show markedly inferior performance or cannot do the task at all. Due to these limitations, explainable AI has been referred to as a "false hope" (19). Therefore, there is a gap between the superficial insight obtained from explainability tools, and the explicit knowledge of the, as yet unknown, diagnostic features that, in principle, contain useful information to support the classification task. In this study, we propose a structured methodology for using explainability to identify the latent information that is used by the AI. Importantly, we utilize explainability tools, not as a precise source of explicit information on how the model makes decisions, but instead as a source of inspiration for forming exploratory hypotheses, which are then validated by statistical tests, eventually leading to scientific discoveries.

The proposed methodology aims to reveal features that are unknown to the human observer, that nevertheless allow a trained CNN to successfully accomplish a classification task. One such task is the classification of patient sex based on retinal fundus images. While invisible to the expert human eye (e.g. ophthalmologist), patient sex is a trait that can be predicted successfully by CNNs based on fundus photographs (12, 27–30). Here, we use discrimination of male vs. female eyes as a case study to validate the efficacy of our methodology. This particular classification task, although limited in clinical utility, serves as an ideal case for a proof-of-concept study considering the reliance of CNNs on large datasets. Indeed, patient sex is a readily available label included in most medical imaging datasets with balanced samples of the two classes.

Although ophthalmologists are not trained to recognize patient sex in retinal imaging, it is conceivable that subtle differences exist between the male and female retinas. Studies using optical coherence tomography (OCT) have suggested morphological differences between male and female retinas, including retinal thickness and retinal nerve fiber layer thickness (31, 32). In addition, sex differences in ocular blood flow have been reported, though few empirical studies have examined this issue [see (33) for a review]. As the first study ever to show the ability of CNNs in predicting patient sex based on fundus images, Poplin et al. (12) used attention maps to highlight the regions that the trained model uses to make the predictions. They showed that for sex classification, the model mainly uses the optic disc and blood vessels, and these anatomical areas are highlighted in attention maps in 78% and 71% of the samples, respectively. Other studies have also suggested the optic disc, macula, and retinal vasculature as possible sources of sex differences relying on the saliency map results obtained from their trained models (29, 34). Another study has used the BagNet model as an interpretable-by-design architecture to classify and explain sex, based on retinal fundus images (28). They demonstrated that the optic disc and macula provide most evidence for males and females, respectively. However, they stated that the specific features and sex differences within these anatomical areas contributing to the prediction are yet to be found. Inspired by AI's performance, Yamashita et al. (35) compared numerous parameters and measurements available in color fundus images between men and women, and they found various significant differences regarding the peripapillary area, optic disc, and retinal vessels. Combined, the statistically detected features achieved a classification accuracy of 77.9%. Although this is an improvement compared to chance level, there is still a large gap between the achieved accuracy of this study and that of the best CNN results. More importantly, multiple parameters (about 40) were tested using the same dataset without accounting for multiple comparisons problem that can lead to false positive results. Moreover, as these parameters were defined based on clinical knowledge, it is unclear if AI uses the same features in fundus images.

The present study is the first of its kind that searches for specific and clear sex differences visible in retinal fundus images that are hypothesized solely based on deep learning and interpretation techniques and that attempts to train ophthalmologists on the newly discovered features.

Overview of the proposed novel methodology

By taking patient sex as a case study, the proposed methodological pipeline, as summarized in Fig. 1, consists of four phases: (i) CNN Development, (ii) Inspiration, (iii) Exploration, and (iv) Verification. In the first phase, *CNN development*, we train a CNN model on sex classification based on fundus images, and the model's generalization performance is assessed to ensure the task is learned properly. In the second phase, *Inspiration*, post hoc deep learning interpretation techniques are utilized to generate visualizations of the model's decision process. These visualizations are used solely for the purpose of making observations and gleaning insight. This stage is equivalent to the early phase of scientific method, where observations in nature are used to produce exploratory hypotheses—here we make observations of the model's decision process. Due to the aforementioned ambiguity in the interpretation of AI explainability tools, a number of exploratory hypotheses are liberally proposed at this stage. These hypotheses consist of specific, quantitative differences in image-based parameters between the two classes (males vs. females). In the third phase, *Exploration*, all exploratory hypotheses are tested on an independent dataset that has not been used in the CNN development stage. In the fourth phase, *Verification*, the hypotheses that yield significant results in the Exploration phase are re-tested and controlled for multiple comparisons on a new independent dataset that has not been used in any prior phase. The significant differences between the two image classes that are replicated in the Verification phase represent our "discovery."

In this study, our ultimate goal is to discover new biomarkers in medical images and train diagnosticians to identify them. The 4-stage methodological pipeline we introduce enables the discovery of new features, but it does not guarantee that the diagnostician will be able to easily learn them and incorporate them into their skill set. To address this question, we undertake a psychophysical study to test human observers, including an expert diagnostician group (a group of ophthalmologists), and a nonexpert group. This protocol includes a training block as well as pretraining and post-training retinal image recognition tasks. Additionally, we examine any potential associations between the observers' ability to learn the retinal biomarkers and their domain-general visual object recognition ability, which we assess separately using a novel object memory task (36).

Preview

In the CNN development phase, we successfully trained a VGG model to classify sex in retinal fundus images. The subsequent Inspiration phase yielded multiple observations regarding sex-related variations in the optic disc and retinal vasculature, leading us to formulate 14 testable exploratory hypotheses. In the Exploration phase, we found significant differences in nine of these hypotheses. Our Verification phase confirmed five of these hypotheses, one of which identified the peripapillary area to be darker in males. The remaining four positive results pertained to vasculature, revealing significantly greater length, higher number of nodes and branches, as well as greater retinal coverage by vessels in the superior temporal quadrant in males. Finally, to assess the teachability of these newly discovered features, we conducted a psychophysical study. A group of ophthalmologists, previously unable to determine sex from retinal fundus demonstrated a substantial improvement in their ability to do so after a brief training session on the newly discovered retinal features distinguishing males from females.

Methods

Fundus image datasets

Our methodology relies on forming three mutually-disjoint datasets: the *CNN development set* (which is further partitioned into training, validation and test sets, as commonly done in neural network training), the *Exploration Set*, and the *Verification Set*.

We used two different sources of retinal fundus images in this study. The first, ODIR (37) is a publicly available fundus dataset containing 7,000 annotated images from 3,500 patients with information regarding the age, sex, and pathological condition of each eye. In this dataset, samples are labeled for DR, glaucoma, cataract, hypertension, AMD, myopia, and other diseases or abnormalities. Images are collected from different hospitals and medical centers and therefore captured by a variety of fundus cameras with different resolutions and angles of view. To prevent the model from using potential sex differences in the prevalence of diseases, we excluded all images with any eye disease or other abnormalities, as well as images with low quality as identified by a trained ophthalmologist (Dr. Ozturan). Annotations within the ODIR dataset incorporate diagnostic labels specific to each eye. To maximize the inclusion of images labeled as "normal," we employed a filtering process targeting individual eyes instead of whole patients. The resulting dataset consists of 3,146 images from 1,991 patients (comprising 1,601 images from the left eye), all of which contributed to the CNN development set. The summary statistics are reported in online supplementary Table S1. The second source for fundus images was the retinal imaging database of Vancouver General Hospital Ophthalmic Imaging Department, named "VCH source" from here on. It comprises a total of 2100 images from 1,167 patients, with exactly half of the images taken from the left eye. These images, which were not labeled, are selected from individuals with healthy eyes, assessed by Dr. Ozturan to exclude the eyes with signs of ocular disease or abnormality as well as images with low quality; 1,600 images (of 800 patients) of the VCH source was labeled as the "VCH phase-1" set, and contributed to the CNN development set. Online supplementary Table S2 represents the summary statistics of the VCH phase-1 dataset. In sum, a total of 4,746 images (ODIR + VCH phase-1) were placed in the CNN development set and thus used to train and assess the CNN model. Both ODIR and VCH phase-1 portions of the development set were randomly partitioned into training, validation, and test sets with approximately 70, 15, and 15% of the aggregate sets, respectively. Care was taken to make sure all images of any given patient remained in the same partition. All images were cropped to get a square image with equal height and width by detecting the circular contour of the fundus images and placing it at the center of the square.

The remaining 500 images from the VCH source were used to populate the *Exploration* set (100 images, Phase 3) and the *Verification* set (400 images, Phase 4). In both sets, exactly half of the male and female images are taken from the right eye. Age did not differ between male and female patients in both sets ($P > 0.2$). The summary statistics are reported in online supplementary Table S3. Ethnicity is not included in the ODIR and DOVS datasets. This is due to the fact that these sets consist of images originally captured for clinical purposes and subsequently used in a retrospective manner.

The study was approved by the UBC Clinical Research Ethics Board, UBC Behavioural Ethics Board, and Vancouver Coastal Health Research Institute. For the behavioral testing protocol, informed consent was obtained in accordance with the principles in

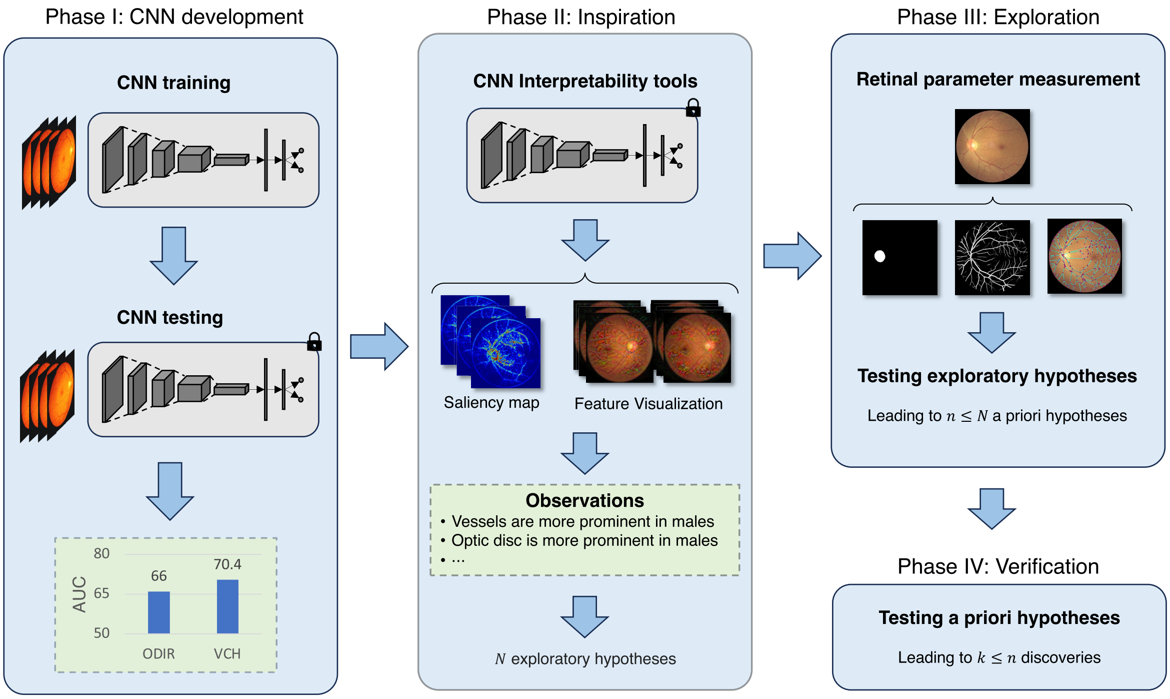
Phase I: CNN development
CNN training
CNN testing
80
65
50
AUC
66
70.4
ODIR
VCH
Phase II: Inspiration
CNN Interpretability tools
Saliency map
Feature Visualization
Observations
• Vessels are more prominent in males
• Optic disc is more prominent in males
• ...
N exploratory hypotheses
Phase III: Exploration
Retinal parameter measurement
Testing exploratory hypotheses
Leading to $n \leq N$ a priori hypotheses
Phase IV: Verification
Testing a priori hypotheses
Leading to $k \leq n$ discoveries

Fig. 1. Overview of our methodology.

the Declaration of Helsinki. For access to retinal fundus images, the requirement of consent was waived.

Phase 1—CNN development: sex classification

CNN architecture

The model architecture used in this study is VGG16 (38). The model's parameters were initialized by the pretrained model weights on the ImageNet dataset (39). Since the pretrained model has 1,000 output classes consistent with the ImageNet classification contest, the model's classifier module, which is the last fully connected layer, was replaced with a new randomly initialized fully connected layer containing 4,096 inputs (the number of output features of VGG model) and two outputs corresponding to male and female classes.

Training procedure

We utilized a transfer learning approach followed by a fine-tuning step to train the network using the training subset of the CNN development set. During the first 2 epochs, the network's weights were frozen, while the new classifier layer was learning the task. This is a common technique to prevent the gradient calculated based on the initial random weights of the classifier layer from changing the network's parameters in a direction that is not meaningful and not necessarily aligned with the task. At the conclusion of the first 2 epochs, as the classifier has learned the task to some extent, we unfroze the network's weights and allowed them to change during the subsequent 100 epochs. Hyperparameters were tuned based on the validation performance and by trying various combinations. A summary of the hyperparameters used for training and evaluating the model can be found in online supplementary Table S4.

Data augmentation and transforms

To further improve the performance of the model, we took advantage of data augmentation and image transforms. During the training process, all images were rotated by a random amount chosen uniformly from −10 to +10 degrees to prevent the network from memorizing image-label pairs. Furthermore, we utilized a novel idea, not used before in similar studies to the best of our knowledge, that can be applied to fundus image datasets specifically because of their nature: The left and right retinas are anatomical mirror-images, approximately symmetrical along vertical axis, leading to a large image-level change (left vs. right) that is not related or informative to the sex classification task. We removed this image variance across the dataset by horizontally flipping all images taken from the right-eye so they appear like a left-eye fundus image. This "horizontal flipping" transform removes part of the image variance across the dataset that is known a priori to be irrelevant to the task and improves the model performance for sex classification. Arguably this is because horizontal flipping allows the model to expect the same anatomical parts of the retina in nearly same locations of the input image (i.e. optic disc on the left side and fovea on the right side) and, in turn, learns the features more efficiently.

Model evaluation

The validation subsets of the CNN development dataset was used for evaluating the Model during the training process and tuning the hyperparameters. At every training epoch, various performance metrics were calculated and recorded on both training and validation sets: area under the receiver operating characteristic curve (AUC) score, accuracy, hit rate, false alarm rate, and binary cross-entropy (BCE) loss. Once the training course is over, the epoch at which the highest validation AUC occurred is selected to report validation metrics. In addition, the Model's weights from the same epoch were saved as "the best model's weights." Then, the best Model was reloaded to obtain the performance metrics separately on the two unseen test sets obtained from ODIR and VCH phase-1 to assess generalizability performance.

We used nonparametric bootstrapping to evaluate the significance of the results. We generated $B = 1, 000$ bootstrap replicates of the test sets to obtain the confidence interval for each performance metric. The chance level calculated from the ratio of male images in each test set was then compared to the AUC confidence intervals. The *P*-values were calculated based on the percentile rank of chance-level (50%) performance in the bootstrap AUC distribution.

In order to maintain a benchmark performance for comparison, two models are trained independently using the identical procedure described in the Training procedure section, namely, "Trained Model," and "Random Model." Random Model was trained on the same datasets with the only difference that male and female labels were randomly shuffled in advance to training. Performance metrics on the test sets are reported for the two models.

Phase 2—Inspiration

We used the Grad-CAM (20) technique to generate saliency maps. In this method, input images were first normalized by the average and SD values of each color channel calculated based on the ImageNet dataset [see online supplementary Table S5], the data on which the pretrained models are trained. Each image was then fed to the network to complete the forward path needed for calculating the gradient during the backward process, and the predicted label was saved. The model's output was one-hot coded, i.e. the output corresponding to the predicted class was set to one and the other class to zero. Next, the Grad-CAM saliency map was generated by back-propagating the gradient of the predicted label to the last convolutional layer. In an independent process, the Guided Backpropagation map was also calculated by the deconvolutional network created as an inverse of the trained model (see Ref. (22) for details) and again setting the predicted class's output to one and the other class's output to zero. Finally, these two matrices were multiplied pixel-wise to form the Guided Grad-CAM saliency maps.

To visualize the preferred stimuli for male and female classes, we used the regularized activation maximization method (24). This technique allows the user to input a sample image and provides a transformed version that maximizes the response at some layer of the network. Here, we input both noise images as well as male and female fundoscopic images from the test sets. Similarly, with the saliency map method, input images are normalized before going through the optimization process, and the model's parameters are fixed. Stochastic gradient descent (SGD) was used as the method to maximize the activation of the output class of interest. We tried different combinations of hyperparameters in the implementation of Ref. (40) and selected the values that led to qualitatively more interpretable results. The hyperparameters used to generate the final results are summarized in online supplementary Table S5. We used the PyTorch implementation by Ref. (40) for both Guided Grad-CAM and regularized feature visualization.

Phase 3—Exploration and Phase 4—verification

Measuring retinal parameters

In order to test the exploratory hypotheses, a wide variety of retinal parameters were measured. Thus, we needed to segment the main anatomical parts of the retina, specifically, the optic disc, retinal vasculature, and the peripapillary area. The segmentation results were then used to quantitatively measure the variables of interest and statistically test the hypothesized sex differences.

The optic disc and the fovea were segmented using the Sefexa software (41) under the supervision of an ophthalmologist (Dr. Ozturan). A sample segmentation output is depicted in Fig. 2. The optic disc mask covers the nonparametric oval-like shape of the optic disc, and the fovea mask is a small dot marking the approximate location of the fovea since a precise location cannot be determined in this imaging modality. To segment the vasculature, we trained the LadderNet model (42), a deep learning architecture specialized for image segmentation, using the DRIVE dataset. The implementation by Ref. (43) was used for training the model on the DRIVE dataset (44) and then applying the trained model to the datasets used in this study. See online supplementary material for the details of the vessel segmentation procedure and the DRIVE dataset.

Based on the segmentation masks, multiple measurements were made to statistically test the hypotheses derived from CNN interpretation results. To characterize the optic disc (OD), we measured the area, sharpness of the edge, and brightness. The area covered by the vessels in different parts of the retina, as well as the number of nodes, number of branches, and the total length of branches in the vessel graph, were measured to characterize vasculature. Also, the radius of the foveal avascular zone (FAZ) was estimated. In order to account for the variation in the angular field of view, measurements were normalized by the distance between the fovea and the center of the optic disc, where appropriate. This is the same normalization method used by ophthalmologists; for instance, to assess the changes in optic disc size, they measure its diameter as a fraction of the distance between the fovea and optic disc.

In the definition of the measurements and variables, G denotes the original fundus image with size $w \times h$ converted to gray-scale by using `cvtColor` function from OpenCV Python library. $G_{x,y}$ denotes the value of the pixel located at column x and row y of the gray-scale fundus image. In the binary masks [for the optic disc, vessels, and field of view (FOV)], each pixel can take the value of either zero (not annotated as the region of interest) or one (annotated as the region of interest). $OD_{x,y}$, $V_{x,y}$, and $FOV_{x,y}$ represent the value of pixel (x,y) of the optic disc mask, the vessel mask, and the FOV mask, respectively.

Center of optic disc. The optic disc center is calculated as the average coordinates of pixels included in the optic disc mask:

$$x_{OD} = \frac{\sum_{x,y} x\, OD_{x,y}}{\sum_{x,y} OD_{x,y}}, \quad y_{OD} = \frac{\sum_{x,y} y\, OD_{x,y}}{\sum_{x,y} OD_{x,y}}$$

Distance between optic disc and fovea. This distance is calculated to normalize length and area measurements to account for any differences in the angular view of the fundus images.

$$d_{OD-F} = \sqrt{(x_{OD} - x_F)^2 + (y_{OD} - y_F)^2}$$

Normalized area of optic disc. The area of the optic disc is calculated as the number of nonzero pixels in the optic disc mask. The resulting area is then normalized by a factor of d^2_{OD-F} as its dimensionality is *pixel*-squared.

$$A_{OD} = \sum_{x,y} OD_{x,y}, \quad \bar{A}_{OD} = \frac{A_{OD}}{d^2_{OD-F}}$$

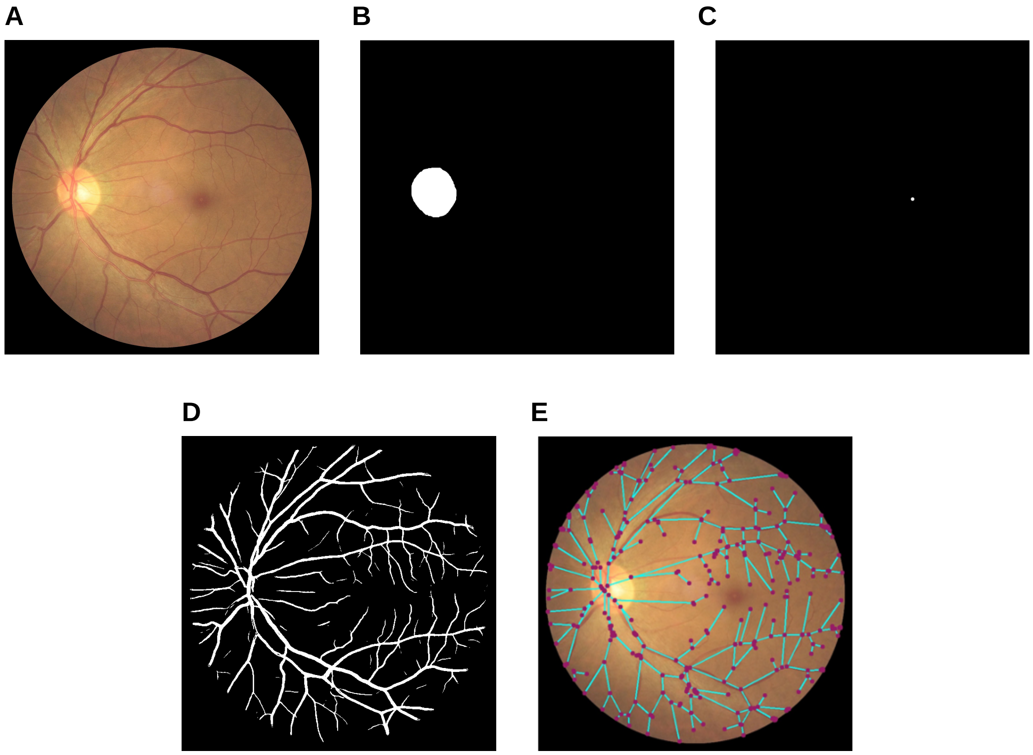

Fig. 2. A sample fundus image A) along with its binary optic disc mask B), fovea C), and vessel mask D). An illustration of the vessel graph extracted based on binary vessel mask is depicted in panel E). Lines and dots represent the edges and nodes of the obtained vessel graph, respectively.

Average normalized brightness of optic disc. The gray-scale image is first normalized in terms of brightness by shifting its average to 0.5.

$$\bar{G} = G - \frac{\sum_{x,y} G_{x,y}}{w \times h} + 0.5$$

This normalized gray-scale image, along with the optic disc mask, is then used to calculate the average brightness of the optic disc (denoted as B_{OD}).

$$B_{OD} = \frac{\sum_{x,y} OD_{x,y}\,\bar{G}_{x,y}}{\sum_{x,y} OD_{x,y}}$$

Sharpness of optic disc edge. The Sobel operator is applied on the optic disc mask to obtain the contour of the optic disc (S_{OD}). Since the mask is binary, the Sobel operator returns a one-pixel wide contour, which is passed through a Gaussian blur filter (of size 7 × 7) to obtain a wider mask around the optic disc edge (denoted as EM). This resulting edge mask has its highest values on the exact edge of the optic disc and decreases as the pixels get farther from the optic disc edge.

$$S_{OD} = \texttt{Sobel}(OD)$$
$$EM = \texttt{GaussianBlur}(S_{OD})$$

Then the gray-scale image is passed through the Sobel operator to obtain the derivative of the image (denoted as S) and is averaged by the edge mask.

$$S = \texttt{Sobel}(\bar{G})$$

$$\text{Sharpness}_{OD} = \frac{\sum_{x,y} EM_{x,y}\,S_{x,y}}{\sum_{x,y} EM_{x,y}}$$

Brightness of peripapillary area. First, the peripapillary mask is calculated by subtracting the optic disc mask from a circular mask with a radius of 1.4 R_{OD} centered at the optic disc center. The resulting peripapillary binary mask denoted as P, is a ring-shaped mask with an average width of 0.4 R_{OD} around the optic disc.

$$\text{Brightness}_{\text{peripapillary}} = \frac{\sum_{x,y} P_{x,y}\,\bar{G}_{x,y}}{\sum_{x,y} P_{x,y}}$$

Vessel coverage. In addition to the entire fundus, we aimed to measure vessel coverage (VC) separately for different quadrants of the retina, namely, superior temporal (ST), superior nasal (SN), inferior temporal (IT), and inferior nasal (IN). Therefore, images are first aligned and rotated with respect to the center of the optic disc in a way that the line connecting the center of the optic disc to the fovea lies exactly horizontally. Also, the optic disc mask, fovea location, FOV mask, and the vessel mask are rotated accordingly and denoted with r superscript.

$$(\text{VC})_{\text{entire fundus}} = \frac{\sum_{x,y} \text{FOV}_{x,y}\,V_{x,y}}{\sum_{x,y} \text{FOV}_{x,y}}$$

For the right eye we have:

$$(\text{VC})_{\text{ST}} = \frac{\sum_{x<x_c,y<y_c} \text{FOV}^r_{x,y}\,V^r_{x,y}}{\sum_{x<x_c,y<y_c} \text{FOV}^r_{x,y}}$$

$$(\text{VC})_{\text{SN}} = \frac{\sum_{x>x_c,y<y_c} \text{FOV}^r_{x,y}\,V^r_{x,y}}{\sum_{x>x_c,y<y_c} \text{FOV}^r_{x,y}}$$

$$(\text{VC})_{\text{IT}} = \frac{\sum_{x<x_c,y>y_c} \text{FOV}^r_{x,y}\,V^r_{x,y}}{\sum_{x<x_c,y>y_c} \text{FOV}^r_{x,y}}$$

$$(\text{VC})_{\text{IN}} = \frac{\sum_{x>x_c,y>y_c} \text{FOV}^r_{x,y}\,V^r_{x,y}}{\sum_{x>x_c,y>y_c} \text{FOV}^r_{x,y}}$$

And for the left eye we have:

$$(\mathrm{VC})_{\mathrm{ST}} = \frac{\sum_{x>x_c,y<y_c} \mathrm{FOV}^r_{x,y}\, V^r_{x,y}}{\sum_{x>x_c,y<y_c} \mathrm{FOV}^r_{x,y}}$$

$$(\mathrm{VC})_{\mathrm{SN}} = \frac{\sum_{x<x_c,y<y_c} \mathrm{FOV}^r_{x,y}\, V^r_{x,y}}{\sum_{x<x_c,y<y_c} \mathrm{FOV}^r_{x,y}}$$

$$(\mathrm{VC})_{\mathrm{IT}} = \frac{\sum_{x>x_c,y>y_c} \mathrm{FOV}^r_{x,y}\, V^r_{x,y}}{\sum_{x>x_c,y>y_c} \mathrm{FOV}^r_{x,y}}$$

$$(\mathrm{VC})_{\mathrm{IN}} = \frac{\sum_{x<x_c,y>y_c} \mathrm{FOV}^r_{x,y}\, V^r_{x,y}}{\sum_{x<x_c,y>y_c} \mathrm{FOV}^r_{x,y}}$$

Macular vessel coverage is also measured using a circular mask centered at the fovea with the radius of $0.5\, d_{\mathrm{OD-F}}$. The macula mask is denoted as M.

$$(\mathrm{VC})_{\mathrm{macula}} = \frac{\sum_{x,y} M_{x,y}\, V_{x,y}}{\sum_{x<x_c,y>y_c} M_{x,y}}$$

FAZ radius. The radius of FAZ is reported as the radius of the largest circle centered at the fovea, which does not contain any vessels according to the binary vessel mask. If we denote a circle mask centered at the fovea with radius r as $C(r)$, we have:

$$\begin{aligned} R_{\mathrm{FAZ}} &= \max \\ &\quad \text{s.t.} \quad \sum_{x,y} C(r)_{x,y}\, V_{x,y} = 0 \\ \bar{R}_{\mathrm{FAZ}} &= \frac{R_{\mathrm{FAZ}}}{d_{\mathrm{OD-F}}} \end{aligned}$$

Vessel graph properties. To further analyze the structural properties of the retinal vasculature, the binary vessel masks were also translated to vessel graphs by a Python library named `Skan` designed for skeleton image analysis. The obtained vessel graph contains structural information such as details of branches and nodes (2). Based on the output of the graph analysis, the number of branches, number of nodes, and the total length of branches for the vessel graph were calculated.

Statistical analysis

A two-sample t-test (male eyes vs. female eyes) has been performed for each of the retinal measurements. For the short-listed hypotheses (Phase-4 Verification), the P-values were then adjusted (45) using the Benjamini-Hochberg method (46) to control for multiple comparisons.

Behavioral testing of ophthalmologists and nonexperts

We examined domain-general object recognition ability and sex-recognition performance based on retinal fundus images in a group of experts and a group of naive observers. Assessment of object recognition ability was based on accuracy on the Novel Object Memory test (NOMT) using the Ziggerins novel object category (36). Sex-recognition task was completed twice, once before, and again after a training block.

Participants

Twenty-six participants in the expert group (all ophthalmologists; 16 females; age: $M = 34.85$, $\mathrm{SD} = 6.70$) and 31 naive participants with no experience regarding fundus images (18 females; age: $M = 34.00$ years, $\mathrm{SD} = 12.26$) took part in the study.

Procedure

Sex-recognition block. Participants completed a 100-trial sex-discrimination block once before, and again after, the training block. At each trial one male and one female fundus image was displayed side-by-side on the screen and the participants were asked to choose the male image in a two-alternative forced-choice (2-AFC) paradigm. The images remained on the screen until a response was made. No feedback was provided. The percentage of correct responses over the entire block was used as the outcome performance measure.

Training block. The training block consisted of didactic and practical components. In the didactic component, participants viewed a short presentation of descriptions of retinal characteristics that have been found to differ between males and females (e.g. brighter peripapillary region in females, greater vascular prominence in the superior temporal quadrant in males) as well as visual illustrations of how these might present in fundoscopic images. In the practical component, participants completed 50 trials of a sex-recognition task in which they chose the male image among a male-female pairs in a 2-AFC paradigm. Feedback was provided which highlighted the correct choice. All participants completed both the didactic and the practical portions in that order, which lasted approximately 15 min.

NOMT block. The methodological details of the NOMT test is described in Richler et al. (36), which we will summarize here. The NOMT is a 72-trial 3-AFC protocol which consists of a 18-trial learning phase and a 54-trial test phase. During the learning phase, participants are familiarized with six distinct exemplars of a novel object category (in our case, the Ziggerins category) by viewing each in three different viewpoints. The six targets are studied for a 20 s period, once at the start of the test phase and again at the mid-point. At each of the 72 trials, participants are asked to select the previously seen exemplar among two distractors. Percentage of correct responses across all trials are used as the outcome performance measure.

Results

Phase 1—CNN development: sex classification

Training metrics

To track the training progress, the network's performance was measured in real-time over the epochs on both training and validation sets. Three evaluation metrics, namely AUC, accuracy, and BCE loss, are plotted separately for the training (blue) and validation (red) sets in online supplementary Fig. S1.

Generalization performance

The model was tested on the unseen ODIR and VCH phase-1 test partitions, alongside a *Random model* that was trained on randomly shuffled female/male labels. To obtain the significance level of the results and calculate P-values, performance metrics achieved by each model were compared to chance level performance (see Table 1).

On the unseen ODIR test partition, the AUC achieved by the CNN model was 0.658, significantly greater than chance level ($Ps < 0.001$). The *Random model* achieved 0.480 AUC ($P = 0.992$), which was not significantly different from (and very close to)

Table 1. Test performance on the unseen test partitions of the ODIR and VCH phase-1 datasets.

	AUC (CI_α)	*P*-value
(a) ODIR		
Trained model	**0.658 (0.611, 0.704)**	**<0.0001**[a]
Random model	0.480 (0.425, 0.533)	0.992
(b) VCH phase-1		
Trained model	**0.728 (0.667, 0.789)**	**<0.0001**[a]
Random model	0.422 (0.351, 0.495)	0.979

AUC values, along with their corresponding confidence intervals and *P*-values, are reported separately for the models on ODIR and DOVS test sets. Performance significantly above chance level shown in bold. [a]Significant *P*-values.

the chance level, as expected. The 95% confidence intervals (CIs) are calculated based on nonparametric bootstrapping with $B = 1,000$ bootstrap resamples of the ODIR test set with size $N = 480$. AUC, *P*-values, and confidence intervals obtained from ODIR by the two models are shown in Table 1, part (a). The results imply that the network classifies patient sex with significantly higher than chance level at an $\alpha = 0.05$ confidence level.

Based on the unseen VCH phase-1 test partition, the CNN model obtained 0.728 AUC ($P < 0.001$). The *Random model* again performed similar to chance level with 0.422 AUC ($P = 0.979$). $B = 1,000$ bootstrap resamples of the test set with $N = 240$ sample size were used to calculate the confidence intervals. The AUC scores, *P*-values, and CIs obtained from VCH phase-1 by the two models reported in Table 1, part (b) indicate that the network has obtained a sex classification performance significantly better than the chance level.

Phase 2—Inspiration

Saliency maps

Figure 3 depicts four sample saliency map results for each of the male and female classes. The original fundus images fed to the model, the Guided Grad-CAM outputs, and the color-coded saliency maps are shown in the left, middle, and right panels, respectively. According to these sample maps, the network appears to be attending mainly to the optic disc, retinal vasculature, and to some extent, the fovea. This pattern is consistent among different samples in both male and female groups, suggesting that the information needed for sex classification based on fundus photographs resides in these anatomical structures of the retina.

Feature visualization

Figure 4 shows four sample feature visualization results initialized with a fundus image (middle column). The initial images are changed by the model to a more male-like and a more female-like fundus in the left and right columns, respectively. Upon reviewing a series of feature visualizations, a consistent observation regarding differences between synthetic male and female images was the presence of many tubular vessel-like structures added to the original image for the male synthetic samples. In contrast, this pattern was not observed for the female synthetic samples. These added tubular structures are relatively thick and similar to the main veins and arteries seen in the original fundus images. This may suggest that from the model's perspective, males have more prominent and thicker retinal vasculature compared to females. A second consistent observation was the optic disc visualized to be more prominent in males, showing sharper edges and more contrast to the background. Unlike males, the optic discs are visualized as diminished in the female synthetic images. These were noted as the most consistent patterns observed over many feature visualization outputs.

Based on the observations on the saliency map and feature visualization results, possible sex differences in the retina were described as general exploratory hypotheses under two main themes: (i) retinal vessels are more prominent in males than females, (ii) optic disc is more prominent in males than females. Deriving from these two themes, 14 specific exploratory hypotheses were proposed. (1) The normalized area of the optic disc is significantly larger in males; (2) the normalized brightness of the optic disc is higher in males; (3) optic disc edge is sharper in males; (4) peripapillary area is darker in males; (5) male eyes have higher vessel coverage in the entire fundus; (6) superior temporal quadrant; (7) inferior temporal quadrant; (8) superior nasal quadrant; (9) inferior nasal quadrant; and (10) macula; (11) foveal avascular zone (FAZ) normalized radius is greater in females; (12) in vessel graphs, number of nodes, (13) number of branches, and (14) total length of branches are greater in males.

Phase 3—Exploration

Of the 14 retinal parameters associated with the above hypotheses that were tested on the exploration dataset, 9 showed significant differences between males and females in the expected direction. These are hypotheses 1, 4, 6, 7, 10–14. The average and standard deviation of all tested parameters for males and females, along with the *P*-values and effect sizes, are reported in Table 2.

Phase 4—Verification

Of the nine parameters that resulted in significant sex differences on the exploration dataset in Phase 3, five again showed significant results in the expected direction controlled for multiple comparisons on the verification dataset. Male fundus images showed more nodes and branches in the vessel graph as well as a greater total length of branches. Also, the area covered by the vessels in the superior temporal quadrant of the retina was higher in men. These results confirm that retinal vasculature is more prominent in males compared to females. However, no significant differences were observed in vessel coverage of the other quadrants of the retina and in the FAZ radius. In addition, results confirm significantly darker peripapillary areas in males. There was no significant difference in the normalized area of the optic disc. The verification results appear in Table 3 in which the average and standard deviation for all parameters for male and female groups, along with the effect sizes, *P*-values, and BH-adjusted *P*-values, are reported.

Behavioral testing of ophthalmologists and nonexperts

Mean accuracy in the domain-general object recognition test [the NOMT task (36)] was $M = 81.4\%$ ($SD = 13.4$) for the expert ophthalmologist group, and $M = 78\%$ ($SD = 10.4$) for the nonexpert group. Performance in both groups closely followed the results in Richler, Jeremy & Gauthier (36), who reported $M = 84.4\%$ ($SD = 11.2$), and did not differ between the two groups [$t(55) = 1.095$, $P = 0.28$].

Pretraining sex-recognition accuracy was $M = 51.62\%$ ($SD = 6.51$) for the expert ophthalmologist group, and $M = 51.97\%$ ($SD = 7.97$) for the nonexpert group. Pretraining performance did not differ from chance level (50%) in the expert ophthalmologist group [$t(25) = 1.27$, $P = 0.22$] and in the nonexpert group

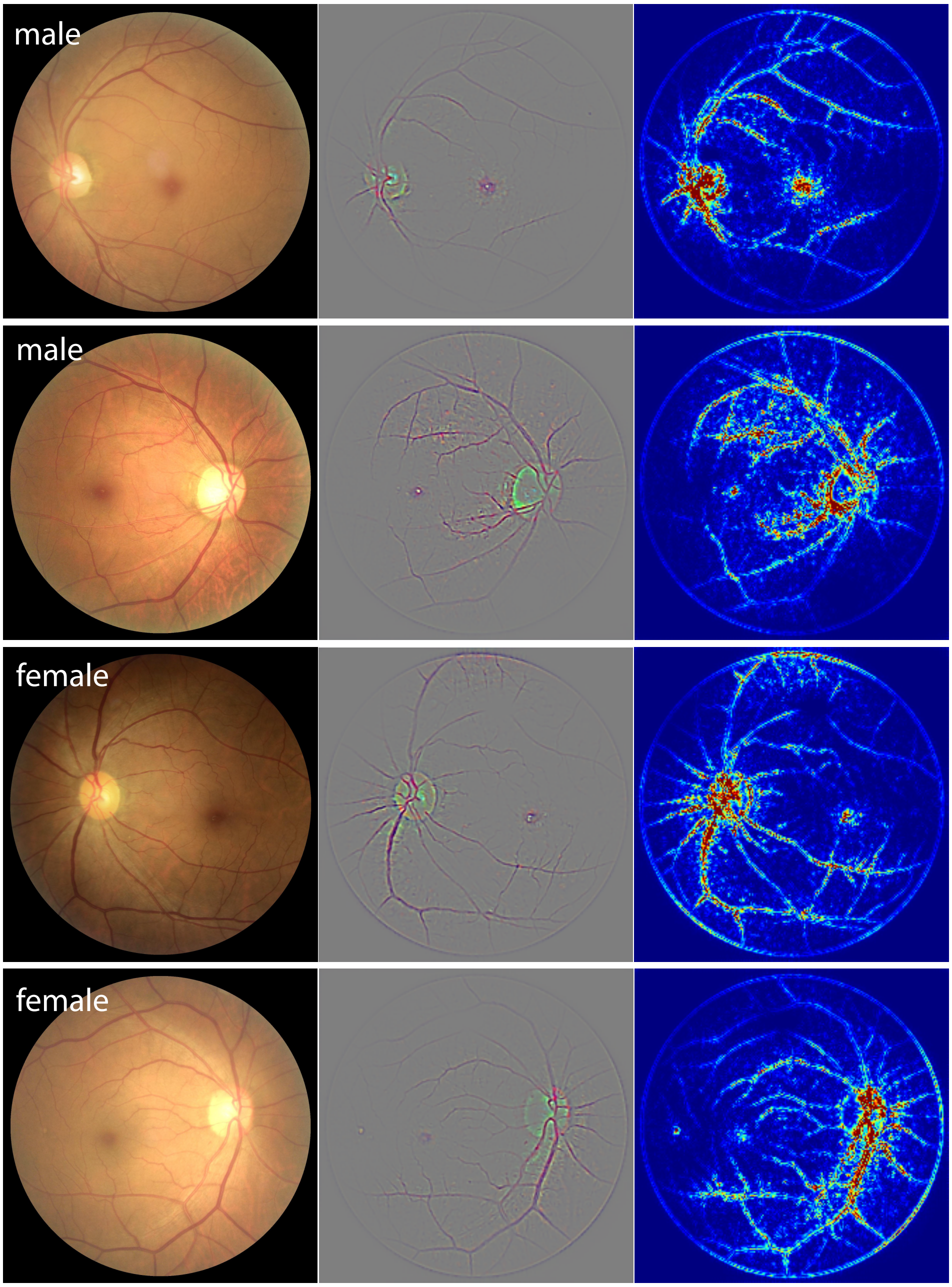

Fig. 3. Saliency map results of sample fundus images from two male and two female patients. In each panel, the original fundus image, the Guided Grad-CAM output (3-channel image), and its color-coded amplitude (single-channel image) are shown from left to right.

[$t(30) = 1.37$, $P = 0.18$], and performance did not differ between the two groups [$t(55) = -0.18$, $P = 0.86$].

Sex-recognition accuracy in the training block was $M = 74.31\%$ ($SD = 10.68$) for the expert ophthalmologist group, and $M = 80.23\%$ ($SD = 10.17$) for the nonexpert group. Training performance was significantly greater than chance level (50%) in the expert ophthalmologist group [$t(25) = 11.60$, $P \ll 0.001$, $d = 2.28$] and in the nonexpert group [$t(30) = 16.57$, $P \ll 0.001$, $d = 2.98$]. In addition, nonexpert group accuracy during training was significantly higher than that of the expert group [$t(55) = -2.14$, $P = 0.036$, $d = 0.59$].

Figure 5 shows pretraining and post-training performance for the two groups. There was a significant increase in sex-recognition accuracy in the expert ophthalmologist group ($M = 65.89\%$, $SD = 4.03$, $P \ll 0.001$, $d = 2.64$), and in the nonexpert group ($M = 66.16\%$, $SD = 3.73$, $P \ll 0.001$, $d = 2.28$). There was no difference in the degree of improvement between the two groups [$t(55) = 0.04$, $P = 0.97$].

Performance in the NOMT was not correlated with degree of post-training improvement in sex recognition accuracy in the expert group ($r = -0.08$, $P = 0.7$) and in the nonexpert group ($r = 0.1$, $P = 0.58$). However, NOMT performance was correlated with performance during the training block for the expert group ($r = 0.6$, $P = 0.001$) and the nonexpert group ($r = 0.39$, $P = 0.028$) (Fig. 6). Interestingly, the correlation in the expert group appears to be largely driven by one outlier participant. When this outlier is removed, the correlation in the expert data is no longer significant ($r = 0.2$, $P = 0.33$).

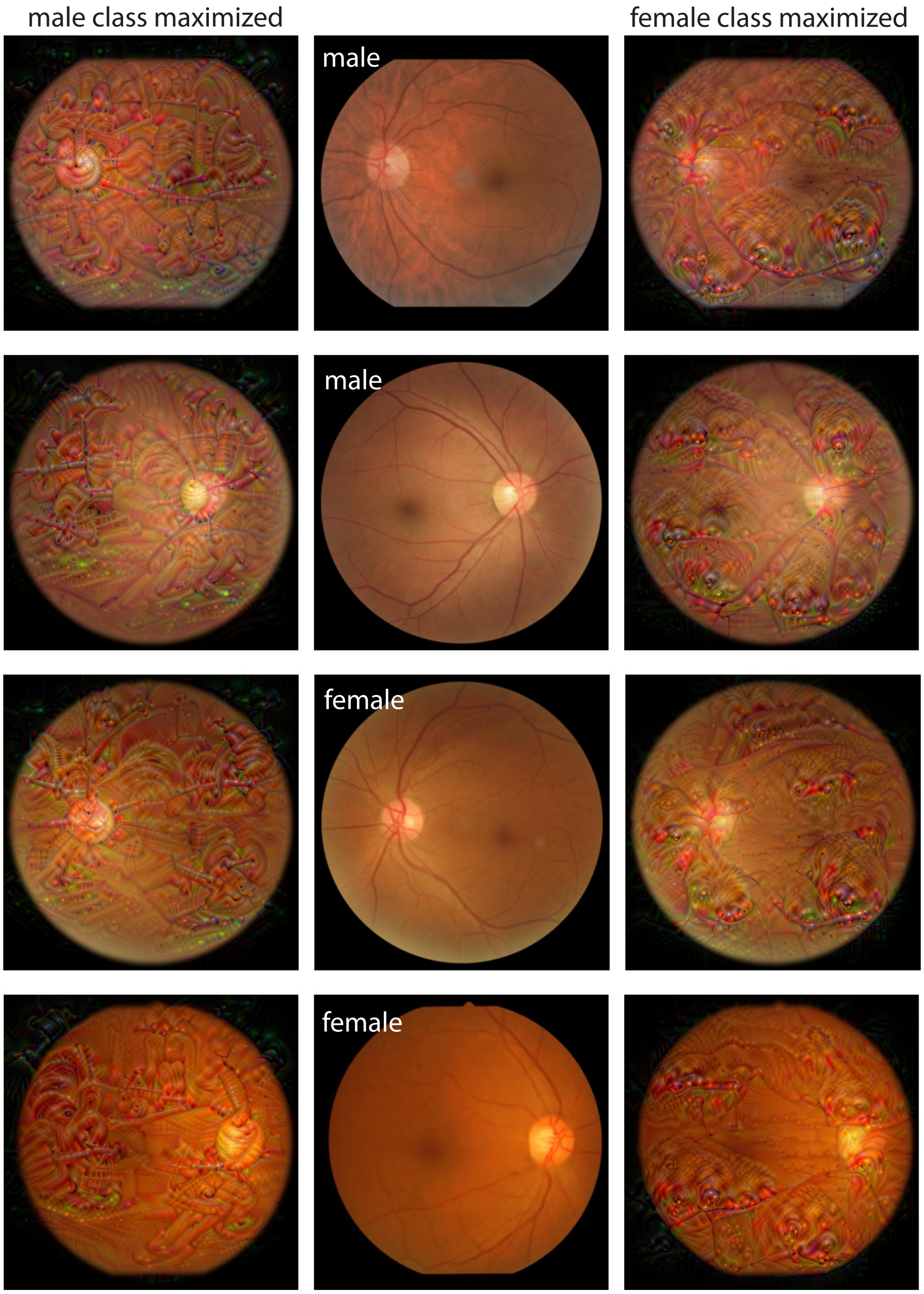

Fig. 4. Feature visualization results for sample fundus images. The top two rows of the middle column show original male images and the bottom two rows show original female images. The left and right columns represent feature visualizations for male and female classes, respectively.

Discussion

The scientific method has evolved and undergone refinement over centuries to study and understand nature. Here, we introduce a minor modification to this method to study and understand AI. Rather than drawing inspiration from natural phenomena and conducting observations in the environment, we derive insights from post hoc interpretation and abstract visualizations of the decision-making processes of the CNN. By employing this modified methodology, we were able to formulate and validate exploratory hypotheses, leading us to discover retinal features that distinguish between males and females in fundus images. The retinal features we have uncovered are not an all-encompassing catalog of sex differences present in the retina. Indeed, our methodology is contingent upon an *Inspiration phase*, which is inherently subjective in nature, involving the identification of recurring patterns, and categorization under overarching themes. Moreover, the method is constrained by the particular interpretation and visualization techniques employed. For instance, previous research has indicated that the macula may potentially exhibit sex differences in the retina [e.g. (12, 28)]. However, this theme was not identified during our Inspiration phase and, consequently, was not explored in the current study. Although our Inspiration phase is inherently ambiguous and subjective, it nevertheless conveyed a sufficiently strong veridical signal,

Table 2. Results for sex differences for the 14 exploratory hypotheses described in Phase 2, tested on the exploration dataset.

Measurement	Male average (SD)	Females average (SD)	Effect size	P-value
OD normalized area	0.117 (0.016)	0.105 (0.020)	0.712	**<0.0001**
OD normalized brightness	0.808 (0.068)	0.813 (0.083)	0.072	0.3610
OD edge sharpness	0.198 (0.019)	0.813 (0.023)	0.223	0.1362
Peripapillary area brightness	0.652 (0.052)	0.670 (0.051)	0.349	**0.0437**
Vessel coverage				
Superior temporal	0.147 (0.014)	0.138 (0.015)	0.606	**0.0017**
Inferior temporal	0.148 (0.016)	0.139 (0.018)	0.513	**0.0063**
Superior nasal	0.190 (0.035)	0.180 (0.034)	0.281	0.0834
Inferior nasal	0.165 (0.038)	0.157 (0.032)	0.227	0.1324
Entire fundus	0.150 (0.013)	0.142 (0.015)	0.590	**0.0022**
Macula	0.127 (0.017)	0.120 (0.020)	0.391	0.0278
Vessel graph properties				
Number of nodes	371.04 (48.35)	326.94 (41.29)	0.981	**<0.0001**
Number of branches	378.62 (49.05)	328.56 (44.16)	1.073	**<0.0001**
Total length	11776.53 (846.25)	11056.87 (978.21)	0.787	**<0.0001**
FAZ normalized radius	0.118 (0.024)	0.131 (0.022)	0.554	**0.0036**

Significant results are shown in bold.

Table 3. Results for sex differences for the nine exploratory hypotheses that showed significant differences in Phase 3, re-tested on the verification dataset.

Measurement	Male average (SD)	Females average (SD)	Effect size	BH-adjusted P-value
OD normalized area	0.110 (0.022)	0.107 (0.022)	0.119	0.1770
Peripapillary area brightness	0.669 (0.054)	0.682 (0.054)	0.243	**0.0234**
Vessel coverage				
Superior temporal	0.142 (0.019)	0.139 (0.017)	0.194	**0.0484**
Inferior temporal	0.142 (0.021)	0.143 (0.019)	0.049	0.4039
Entire fundus	0.145 (0.018)	0.145 (0.017)	0.007	0.4160
Vessel graph				
Number of nodes	349.31 (57.51)	332.66 (53.70)	0.299	**0.0136**
Number of branches	351.44 (62.82)	334.98 (58.19)	0.272	**0.0157**
Total length	11381.20 (1173.34)	11147.80 (1085.52)	0.206	**0.0451**
FAZ normalized radius	0.1295 (0.032)	0.1288 (0.032)	0.021	0.4160

Significant results are shown in bold.

enabling ophthalmologists to receive training on the identified features. Following this training, for the first time, these clinicians were able to distinguish the sex of individuals from fundus images.

Our approach is not without precedence—there has been a growing emphasis on the use of the scientific method in recent discussions surrounding the understanding of AI behavior. Some prominent examples include Olah and colleagues' natural science approach (47) to investigating the behavior and inner workings of artificial neural network models, and Miller's suggestion of incorporating existing body of research in philosophy and social science to advance explainable AI (48). Our approach is broadly aligned with these ideas in so far as focusing on saliency maps and feature visualizations as objects of scientific inquiry. The main divergence of our method is in the Inspiration phase where we introduce a human observer element that generates a list of exploratory hypotheses. A similar human-introduced influence is evident, albeit to a lesser extent, in the work of Bau and colleagues (49). In their analytic method, termed network dissection, Bau et al. (49) identify object-detector units in a CNN, tuned to concepts that were not part of the set of labels the model was trained on. The introduction of the list of concepts is, in a manner, analogous to the human generated list of exploratory hypotheses in the present work, though arguably significantly less subjective in nature compared to our methodology. Nevertheless, it is worth mentioning that all natural science starts with human observers (i.e. scientists) making purely subjective observations in nature and generating exploratory hypotheses, and therefore, we argue that this aspect of our methodology is not a defect, but a feature of our method.

In this study, we used five distinct and nonoverlapping datasets of retinal fundus images. Of the five datasets, three were utilized in the CNN development stage (consisting of training, validation, and testing sets). The remaining two datasets were reserved for hypothesis testing and validation, referred to as the Exploration and Verification datasets, respectively. It is crucial to maintain the separation between these datasets as it serves to mitigate overfitting and performance inflation during the CNN development phase, and to reduce spurious detections and false alarms in the Exploration and Verification stages. The Exploration set was intentionally selected to be of a small size, thereby allowing only the largest effects to be identified (alongside false alarms), while the Verification dataset was chosen to be 4 times larger, in order to optimize the probability of reproducing any bona fide effects that were inferred during the Exploration phase.

Several CNN models have been trained to achieve near-perfect classification of patient sex using fundus images (12, 28, 29, 34), proving that the signal that allows this is present in retinal fundus images, even though it has gone unnoticed by medical practitioners thus far. CNN performance relies heavily on the amount of data used in development, and the moderate performance we achieved in the present study should be viewed in the context of the very small dataset we used for this purpose. The AUC scores obtained by previous works are compared with the current study

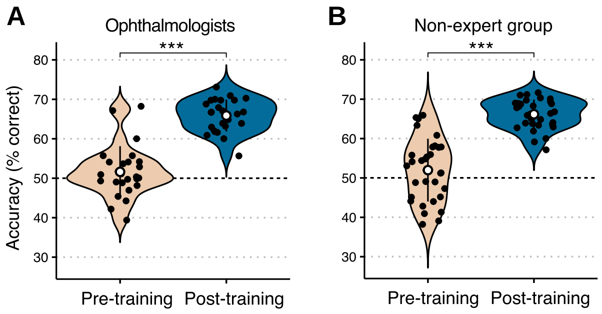

Fig. 5. Accuracy in the 2-AFC sex-recognition task is shown for the pretraining and post-training blocks for the expert ophthalmologist group A) and the nonexpert group B).

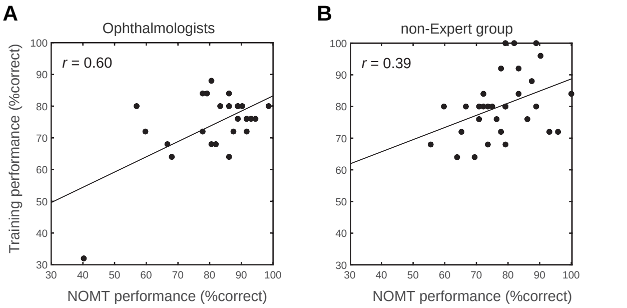

Fig. 6. Accuracy in the training block is shown as a function of NOMT performance for the expert ophthalmologist group A) and the nonexpert group B).

in Table 4. Compared to the massive dataset used in Poplin et al. (12) with about 1.8 million images, our dataset is approximately 540 times smaller, yet there is less than 20% reduction in the AUC score achieved in the present study. A study by Berk et al. (27) provides a more commensurate comparison as they used similar datasets with similar sizes and obtained a test AUC score of 0.72. It is worth mentioning that we deployed a simple architecture and classification paradigm for the purpose of explainability whereas Berk et al. (27) boosted their classification score by ensembling ten separately trained networks with median AUC of 0.69.

Although the CNN development dataset used in the present study is small, and consequently, the generalization performance is modest, the results are nevertheless significantly above chance level. Critically, the features revealed here allowed ophthalmologists and nonexpert observers to learn to recognize patient sex in retinal fundus images. Similarly, post-training recognition of our human participants also represent a modest increase over pretraining levels and chance-level performance. This is not surprising. Indeed, performance of the human observers are on par with the performance of the CNN, consistent with the idea that the human observers were able to adopt all the knowledge leveraged from the CNN. It is also important to note that the clinical utility of our methodology has yet to be proven. Nevertheless, the potential applications of our proposed methodology extend beyond determining patient sex. It could be expanded to identify other clinically significant traits and features, including discovery of retinal biomarkers for ocular, systemic and neurodegenerative diseases, such as AD.

One interesting aspect of this work was the finding that our nonexpert observers performed on par with the ophthalmologists, both pretraining and post-training. This indicated that the years of experience reviewing fundus images to diagnose a variety of ocular diseases did not confer any advantages to the ophthalmologists in the learning of this new task. Work by Biederman and Shiffrar (50) on sexing of day-old chicks provides some context to these pattern of results. This particular task was regarded as highly challenging and requiring substantial experience and expertise. However, when a straightforward distinguishing feature was identified and explained to nonexpert observers, they were able to quickly learn the task and perform on par with experts (professional sexers with many years of experience). This suggests that the primary difficulty may lie in identifying the crucial feature and rule, rather than in the process of learning and applying the rule. Once directly pointed out, both novice and experienced observers can readily adopt the diagnostic strategy. This offers a potential explanation to why our nonexpert controls performed similarly with our ophthalmologists.

Table 4. Sex classification results of the previous studies and the current study.

	Training set images	AUC	CI$_\alpha$
Poplin et al. (12)	1,779,020	0.97	(0.96, 0.98)
Korot et al. (29)	173,819	0.93	—
Current work	**3,306**	**0.728**	**(0.667, 0.789)**
Berk et al. (27)	1,746	0.72	(0.67, 0.77)

Performance significantly above chance level shown in bold.

Interpretation of medical images for the purpose of diagnosis is, in part, a visual recognition task. Diagnosticians vary in their performance at medical image interpretation [e.g. (51)], and although some of this variability is presumably based on experience, it may also be partly due to variation in inherent visual perceptual abilities [e.g. (52)]. To examine this possibility, we measured performance in NOMT to assess a domain-general visual ability and examined its association with our human observers' improvement in post-training sex-recognition scores. We did not find any association between NOMT scores and post-training sex-recognition scores, however, we found a significant positive correlation between NOMT scores and performance during the training block. It is unclear why this pattern of results is observed, though it is possible that the task is learned in a relatively short amount of time by the end of the training block, and no learning effects persist into the post-training block. Smithson et al. (53) asked novice observers to classify white blood cells as cancerous vs. noncancerous, and found a positive correlation between their performance in this task and their domain-general visual object recognition ability. However, this association was only present when trial-by-trial feedback was provided. No correlation was found when feedback was not provided in all trials. Our findings are consistent with this pattern of results: solely in the training block, where trial-by-trial feedback was provided, a significant correlation between retinal sex-recognition and NOMT performance was found. This suggests that visual object recognition ability may be primarily linked to visual learning ability. As learning without feedback is limited, this association may not persist in such a context.

Conclusions

Our study demonstrates the potential of AI to identify and classify traits that are currently not recognized in retinal fundus images.

Using patient sex as a case study, we developed a methodology to leverage the trained AI system to discover new retinal features that differentiate between male and female eyes. Notably, ophthalmologists who were previously unaware of sex-related signs in the eye were trained on the newly discovered retinal features, and were able to differentiate patient sex, indicating the potential for clinical translation of AI-driven discoveries. This proof-of-concept study is the first of its kind to leverage new knowledge from an AI system and sets the stage for future research to explore the potential for clinical translation for various conditions that may not be detectable in the eye through conventional diagnostic techniques. Our findings suggest that AI has the potential to revolutionize clinical practice by expanding the range of traits and conditions that can be diagnosed through retinal imaging.

Acknowledgments

This research was supported in part through computational resources and services provided by Advanced Research Computing at the University of British Columbia. The authors also thank the anonymous reviewers for their valuable suggestions. Parts of this manuscript were posted on a preprint: https://arxiv.org/abs/2301.06675

Supplementary material

Supplementary material is available at *PNAS Nexus* online.

Funding

This work was supported by a Natural Sciences and Engineering Research Council of Canada Discovery Grants RGPIN-2019-05554 (I.O.) and an Accelerator Supplement RGPAS-2019-00026 (I.O.), a Canada Foundation for Innovation, John R. Evans Leaders Fund (I.O.), UBC Data Science Institute postdoctoral award (G.O., I.O., & O.Y.), UBC Faculty of Medicine SSRP Award (L.Y.), UBC DMCBH Kickstart award (I.O. & O.Y.), and a UBC Health Innovation Funding Investment award (I.O. & O.Y.), an NSERC Discovery Grant (22R82411) (O.Y.). O.Y. also acknowledges support by the Pacific Institute for the Mathematical Sciences (PIMS). This research was also supported in part through computational resources and services provided by Advanced Research Computing at the University of British Columbia (54).

Author contributions

P.D. carried out CNN model development, segmentation, and extraction of retinal parameters in fundus images, analyzed and interpreted data, and wrote the paper; G.O. curated and compiled retinal data, recruited participants, and collected and analyzed behavioral data, L.Y. prepared behavioral experiments, recruited participants, and collected data, O.Y. conceived and supervised the study, edited the paper, and provided resources; I.O. conceived and designed the study, analyzed the data, wrote the paper, provided resources, and supervised the study. All authors agreed to the final version of the paper.

Previous presentation

These results were previously presented at [ARVO 2023].

Preprints

A preprint of this article is published at [DOI].

Data availability

The trained model's outputs, the ground-truth labels, and the measured retinal parameters used to support the analyses of this study are available at https://github.com/parsadlr/fundus-sex. The ODIR dataset is available for download at https://odir2019.grand-challenge.org/dataset/. Retinal images sourced from Vancouver Coastal Health cannot be shared publicly due to patient confidentiality constraints. Access to these data can be requested by contacting Sasha Pavlovich, Director of Data Access and Governance, Vancouver Coastal Health, at sasha.pavlovich@vch.ca.

References

1 Baraniuk R, Donoho D, Gavish M. 2020. The science of deep learning. *Proc Natl Acad Sci USA*. 117(48):30029–30032.

2 Elul Y, Rosenberg AA, Schuster A, Bronstein AM, Yaniv Y. 2021. Meeting the unmet needs of clinicians from AI systems showcased for cardiology with deep-learning-based ECG analysis. *Proc Natl Acad Sci USA*. 118(24):e2020620118.

3 Esteva A, *et al.* 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 542(7639):115–118.

4 Shen D, Wu G, Suk H-I. 2017. Deep learning in medical image analysis. *Annu Rev Biomed Eng*. 19:221–248.

5 Suzuki K. 2017. Overview of deep learning in medical imaging. *Radiol Phys Technol*. 10(3):257–273.

6 Date RC, Jesudasen SJ, Weng CY. 2019. Applications of deep learning and artificial intelligence in retina. *Int Ophthalmol Clin*. 59(1):39–57.

7 Schmidt-Erfurth U, Sadeghipour A, Gerendas BS, Waldstein SM, Bogunović H. 2018. Artificial intelligence in retina. *Prog Retin Eye Res*. 67:1–29.

8 Gulshan V, *et al.* 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 316(22):2402–2410.

9 Ganjdanesh A, *et al.* 2022. LONGL-Net: temporal correlation structure guided deep learning model to predict longitudinal age-related macular degeneration severity. *PNAS Nexus*. 1(1):pgab003.

10 Peng Y, *et al.* 2019. Deepseenet: a deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs. *Ophthalmology*. 126(4):565–575.

11 Li Z, *et al.* 2018. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology*. 125(8):1199–1206.

12 Poplin R, *et al.* 2018. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng*. 2(3):158–164.

13 Rim TH, *et al.* 2020. Prediction of systemic biomarkers from retinal photographs: development and validation of deep-learning algorithms. *Lancet Digit Health*. 2(10):e526–e536.

14 Alber J, *et al.* 2020. Developing retinal biomarkers for the earliest stages of Alzheimer's disease: what we know, what we don't, and how to move forward. *Alzheimer's Dement*. 16(1):229–243.

15 Lee S, *et al.* 2020. Amyloid beta immunoreactivity in the retinal ganglion cell layer of the Alzheimer's eye. *Front Neurosci*. 14:758.

16 Liao H, Zhu Z, Peng Y. 2018. Potential utility of retinal imaging for Alzheimer's disease: a review. *Front Aging Neurosci*. 10:188.

17 Mirzaei N, *et al.* 2020. Alzheimer's retinopathy: seeing disease in the eyes. *Front Neurosci*. 14:921.

18 Sidiqi A, *et al.* 2020. In vivo retinal fluorescence imaging with curcumin in an Alzheimer mouse model. *Front Neurosci*. 14:713.

19 Ghassemi M, Oakden-Rayner L, Beam AL. 2021. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*. 3(11):e745–e750.

20 Selvaraju RR, *et al.* 2017. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy. p. 618–626.

21 Simonyan K, Vedaldi A, Zisserman A. 2013. Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv 6034. https://doi.org/10.48550/arXiv.1312.6034, preprint: not peer reviewed.

22 Zeiler MD, Fergus R. 2014. Visualizing and understanding convolutional networks. In: European Conference on Computer Vision. Zurich, Switzerland: Springer. p. 818–833.

23 Olah C, Mordvintsev A, Schubert L. 2017. Feature visualization. *Distill*. 2(11):e7.

24 Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H. 2015. Understanding neural networks through deep visualization. arXiv 06579. https://doi.org/10.48550/arXiv.1506.06579, preprint: not peer reviewed.

25 Borowski J, *et al.* 2020. Exemplary natural images explain CNN activations better than state-of-the-art feature visualization. arXiv 12606. https://doi.org/10.48550/arXiv.2010.12606, preprint: not peer reviewed.

26 Zimmermann RS, *et al.* 2021. How well do feature visualizations support causal understanding of CNN activations? *Adv Neural Inf Process Syst*. 34:11730–11744.

27 Berk A, *et al.* 2022. Learning from few examples: classifying sex from retinal images via deep learning. arXiv 09624. https://doi.org/10.48550/arXiv.2207.09624, preprint: not peer reviewed.

28 Ilanchezian I, *et al.* 2021. Interpretable gender classification from retinal fundus images using BagNets. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Strasbourg, France: Springer. p. 477–487.

29 Korot E, *et al.* 2021. Predicting sex from retinal fundus photographs using automated deep learning. *Sci Rep*. 11(1):1–8.

30 Molnar C. 2019. Interpretable machine learning. https://christophm.github.io/interpretable-ml-book/.

31 Lamparter J, *et al.* 2018. Association of ocular, cardiovascular, morphometric and lifestyle parameters with retinal nerve fibre layer thickness. *PLoS ONE*. 13(5):e0197682.

32 Ooto S, Hangai M, Yoshimura N. 2015. Effects of sex and age on the normal retinal and choroidal structures on optical coherence tomography. *Curr Eye Res*. 40(2):213–225.

33 Schmidl D, Schmetterer L, Garhöfer G, Popa-Cherecheanu A. 2015. Gender differences in ocular blood flow. *Curr Eye Res*. 40(2):201–212.

34 Dieck S, *et al.* 2020. Factors in color fundus photographs that can be used by humans to determine sex of individuals. *Transl Vis Sci Technol*. 9(7):8–8.

35 Yamashita T, *et al.* 2020. Factors in color fundus photographs that can be used by humans to determine sex of individuals. *Transl Vis Sci Technol*. 9(2):4–4.

36 Richler JJ, Wilmer JB, Gauthier I. 2017. General object recognition is specific: evidence from novel and familiar objects. *Cognition*. 166:42–55.

37 Shanggong Medical Technology Co Ltd. 2019. Peking University international competition on ocular disease intelligent recognition. https://odir2019.grand-challenge.org.

38 Simonyan K, Zisserman A. 2014. Very deep convolutional networks for large-scale image recognition. arXiv 1556. https://doi.org/10.48550/arXiv.1409.1556, preprint: not peer reviewed.

39 Deng J, *et al.* 2009. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami (FL): IEEE. p. 248–255.

40 Ozbulak U. 2019. Pytorch CNN visualizations. https://github.com/utkuozbulak/pytorch-cnn-visualizations.

41 Fexa A. 2014. Sefexa image segmentation tool (version 1.2.2.5). http://www.fexovi.com/sefexa.html.

42 Zhuang J. 2018. LadderNet: multi-path networks based on U-net for medical image segmentation. arXiv 07810. https://doi.org/10.48550/arXiv.1810.07810, preprint: not peer reviewed.

43 Lee-Zq. 2021. Vesselseg-pytorch: retinal vessel segmentation toolkit based on pytorch. https://github.com/lee-zq/VesselSeg-Pytorch.

44 Staal JJ, Abramoff MD, Niemeijer M, Viergever MA, van Ginneken B. 2004. Ridge based vessel segmentation in color images of the retina. *IEEE Trans Med Imaging*. 23(4):501–509.

45 Benjamini Y, Heller R, Yekutieli D. 2009. Selective inference in complex research. *Philos Trans R Soc A*. 367(1906):4255–4271.

46 Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B (Methodol)*. 57(1):289–300.

47 Olah C, *et al.* 2020. Zoom in: an introduction to circuits. *Distill*. 5(3):e00024-001.

48 Miller T. 2019. Explanation in artificial intelligence: insights from the social sciences. *Artif Intell*. 267:1–38.

49 Bau D, *et al.* 2020. Understanding the role of individual units in a deep neural network. *Proc Natl Acad Sci USA*. 117(48):30071–30078.

50 Biederman I, Shiffrar MM. 1987. Sexing day-old chicks: a case study and expert systems analysis of a difficult perceptual-learning task. *J Exp Psychol Learn Mem Cogn*. 13(4):640–645.

51 Itani M, Assaker R, Moshiri M, Dubinsky TJ, Dighe MK. 2019. Inter-observer variability in the American College of Radiology Thyroid Imaging Reporting and data System: in-depth analysis and areas for improvement. *Ultrasound Med Biol*. 45(2):461–470.

52 Sunday MA, Donnelly E, Gauthier I. 2018. Both fluid intelligence and visual object recognition ability relate to nodule detection in chest radiographs. *Appl Cogn Psychol*. 32(6):755–762.

53 Smithson CJR, Eichbaum QG, Gauthier I. 2023. Object recognition ability predicts category learning with medical images. *Cogn Res Princ Implic*. 8(1):1–10.

54 UBC Advanced Research Computing. 2019. UBC ARC sockeye. https://doi.org/10.14288/SOCKEYE.