tvst

Article

Classifying Retinal Degeneration in Histological Sections Using Deep Learning

Daniel Al Mouiee[1–3], Erik Meijering[1,2], Michael Kalloniatis[4], Lisa Nivison-Smith[4], Richard A. Williams[5], David A. X. Nayagam[5,6], Thomas C. Spencer[6,7], Chi D. Luu[8,9], Ceara McGowan[6], Stephanie B. Epp[6], and Mohit N. Shivdasani[1,6]

[1] Graduate School of Biomedical Engineering, University of New South Wales, Kensington, NSW, Australia
[2] School of Computer Science and Engineering, University of New South Wales, Kensington, NSW, Australia
[3] School of Biotechnology and Biomolecular Science, University of New South Wales, Kensington, NSW, Australia
[4] School of Optometry and Vision Sciences, University of New South Wales, Kensington, NSW, Australia
[5] Department of Pathology, University of Melbourne, Parkville, VIC, Australia
[6] The Bionics Institute of Australia, East Melbourne, VIC, Australia
[7] Department of Biomedical Engineering, University of Melbourne, Parkville, VIC, Australia
[8] Ophthalmology, Department of Surgery, University of Melbourne, Parkville, VIC, Australia
[9] Centre for Eye Research Australia, Royal Victorian Eye & Ear Hospital, East Melbourne, VIC, Australia

Correspondence: Mohit N. Shivdasani, Graduate School of Biomedical Engineering, Room 515A, Samuels Building, University of New South Wales, Kensington, NSW 2033, Australia.
e-mail: m.shivdasani@unsw.edu.au

Received: November 17, 2020
Accepted: April 27, 2021
Published: June 10, 2021

Keywords: deep learning; machine learning; convolutional neural network; retinal degeneration; hematoxylin and eosin (H&E) staining; retinal dystrophy; retinitis pigmentosa

Citation: Al Mouiee D, Meijering E, Kalloniatis M, Nivison-Smith L, Williams RA, Nayagam DAX, Spencer TC, Luu CD, McGowan C, Epp SB, Shivdasani MN. Classifying retinal degeneration in histological sections using deep learning. Transl Vis Sci Technol. 2021;10(7):9, https://doi.org/10.1167/tvst.10.7.9

Purpose: Artificial intelligence (AI) techniques are increasingly being used to classify retinal diseases. In this study we investigated the ability of a convolutional neural network (CNN) in categorizing histological images into different classes of retinal degeneration.

Methods: Images were obtained from a chemically induced feline model of monocular retinal dystrophy and split into training and testing sets. The training set was graded for the level of retinal degeneration and used to train various CNN architectures. The testing set was evaluated through the best architecture and graded by six observers. Comparisons between model and observer classifications, and interobserver variability were measured. Finally, the effects of using less training images or images containing half the presentable context were investigated.

Results: The best model gave weighted-F1 scores in the range 85% to 90%. Cohen kappa scores reached up to 0.86, indicating high agreement between the model and observers. Interobserver variability was consistent with the model-observer variability in the model's ability to match predictions with the observers. Image context restriction resulted in model performance reduction by up to 6% and at least one training set size resulted in a model performance reduction of 10% compared to the original size.

Conclusions: Detecting the presence and severity of up to three classes of retinal degeneration in histological data can be reliably achieved with a deep learning classifier.

Translational Relevance: This work lays the foundations for future AI models which could aid in the evaluation of more intricate changes occurring in retinal degeneration, particularly in other types of clinically derived image data.

Introduction

The methodological approaches to characterizing the anatomy and physiology of the normal and diseased retina are heavily centered around techniques and technologies which can produce and analyze images of this tissue and its components with a high resolution. At a cellular level, histological techniques are widely used to visualize and characterize anatomical, neurochemical, and metabolic characteristics of the retina. The cellular complexity of the retina, however, means analyzing histological data is often very resource intensive, requiring significant time and expertise to accurately interpret images. This issue is particularly prevalent in the assessment of retinal degeneration where histology can reveal a myriad of disease changes including cell loss, functional synaptic changes, corrupted retinal circuitry, cell migration, formation of scar tissue, and even cellular reprogramming.[1–9] Detailed analyses of such changes can play an important role in assessing the preclinical therapeutic or protective effects or even the safety aspects of various treatments such as retinal prosthetics, gene therapy, or optogenetic techniques on retinal neurons,[10] or aid in further understanding treatment mechanisms and predicting the efficacy of specific treatments. Analysis of retinal histology may also be useful in a clinical setting such as for intraocular tumors or uveitis, which requires examination of the underlying pathology to aid in accurate diagnosis.[11,12] For such investigations, through vitreoretinal surgery, retinal biopsies may be performed[13] and histological analysis may be conducted by a retinal pathologist.[14] Thus, there is a need for more efficient ways to assess complex retinal histology images.

Artificial intelligence (AI), and more specifically deep learning, may offer a solution to time consuming, subjective analysis of retinal histology. AI has already been successfully implemented in several areas of ophthalmology, such as analysis of clinical retinal images, including grading of retinal fundus photographs for diabetic retinopathy,[15,16] assessing retinal optical coherence tomography (OCT) images for age-related macular degeneration prognosis,[17–19] and detecting macular holes.[20] With regards to histological data, AI has been extensively used to study histological changes occurring in other organs[21–24] and has also been used in one study to perform a connectomic reconstruction in the inner retina using electron microscopy images.[25] However, to the best of our knowledge, no study has used AI to categorize features of standard hematoxylin and eosin (H&E) based retinal histology images comprising all layers.

Thus, as a first step towards assessing the viability of using AI techniques for analyzing retinal histology, we aimed to develop an automated neural network to perform classification of histological retinal images into predefined stages of retinal degeneration. As the stages of retinal degeneration are based off a set of rules that are not necessarily distinct, but instead represent a continuum of neuronal changes across the degeneration stages,[26] we first attempted to automate the classification of retinal degeneration based on predefined specialist-approved criteria, specific to retinal features clearly observable in H&E images. This included the development and optimization of a convolutional neural network (CNN) capable of classifying H&E-stained retinal sections into different retinal degeneration stages and comparing its performance to classifications made by several observers with varying degrees of expertise in retinal network identification.

Methods

Data Preparation and Preprocessing

The images used for this study were obtained from animals that were part of another study in our group, which assessed the response of visual cortex neurons to electrical stimulation of healthy versus degenerate retina.[27,28] All animal procedures were approved by the Bionics Institute Animal Research Ethics Committee (Project #14 304AB), complied with the Association for Research in Vision and Ophthalmology statement for use of animals in ophthalmic research, and were in accordance with the Australian Code for the Care and Use of Animals for Scientific Purposes and with the National Institutes of Health, USA, guidelines regarding the care and use of animals for experimental procedures.

Adenosine triphosphate (ATP) was used to induce photoreceptor loss followed by retinal degeneration and remodeling in one eye of four healthy adult cats.[27,28] After up to 23 weeks postinjection, the animal was euthanized, transcardially perfused, and the eyes enucleated and the retina dissected. Representative retinal strips (∼2 mm wide) were dehydrated and embedded in paraffin, sectioned at 5 µm, and stained with H&E. Full details regarding the histological processing techniques have been previously published.[29–31] The length of the full retinal sections ranged between 8500 to 9500 µm. These full sections (Fig. 1) were imaged using a high-resolution slide scanner (Aperio Scanscope XT, Leica Biosystems), which used 20×/0.75 Plan Apo objective lenses, 40× scanning magnification, 0.25 µm/pixel resolution, and produced the final files in SVS (TIFF) format.

The data preprocessing procedure involved cropping the sections into images 250 µm in length and only including the retinal portion (i.e., from the retinal pigment epithelium to the nerve fiber layer). We also chose to focus only on the areas of the retinal sections that were not above the suprachoroidally implanted electrode arrays for retinal stimulation, so that any possible confounding damage to the retina caused by

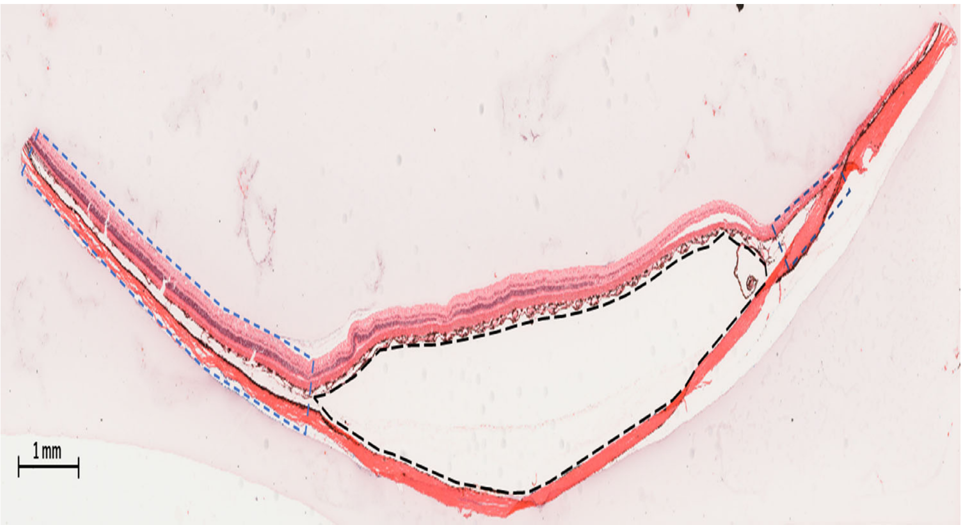

Figure 1. An example of an H&E-stained retinal section. The region enclosed by the black dashed boundary represents the pocket in which electrode arrays were implanted in the suprachoroidal space for another study[26]. The flanking regions enclosed by the blue dashed boundaries represent the retinal segments from which the images were sampled.

acute array implantation was minimized (Fig. 1). We also only selected retinal sections that were largely devoid of major histological artifacts as identified by an expert in retinal networks (MK). After applying the exclusion criteria, 70 sections were analyzed in total; 32 control sections (eight sections from each of the four control noninjected eyes) and 38 degenerated sections (8–11 sections from each of the four ATP-injected eyes). From these 70 sections, 454 cropped images were obtained (Fig. 2). The QuPath open-source software was used for the image cropping and file generation, in conjunction with ImageJ.[32,33]

Degeneration Criteria

Prior to cropping sections, a few samples were first examined by the retinal networks expert (MK) who inferred that it was possible to observe damage gradation along a retinal section. The expert also asserted that distinct biological features could be derived for four different stages of degeneration in the ATP-induced eyes (Table 1). These features were determined using a combination of predefined rules and numerical measurements of some retinal components, as the expert depended on his experience and retinal knowledge to establish such criteria. A discriminative boundary was thus drawn between healthy and diseased retinae, where the healthy instances consistently showed no biological abnormalities (with the exception of some minor remaining histological artifacts such as focal retinal detachments) in any of the cellular layers and the outer segments (OS) of the photoreceptors displayed a rigid formation along the retinal section (Fig. 2a). The inner and outer nuclear layers (INL, ONL) had approximately 3 to 6 and 8 to 15 rows of cell nuclei, respectively. These layers were mostly in parallel alignment and the layers of nuclei within them were closely spaced together making a clearly distinct layer. The retinal pigment epithelium (RPE) was also intact. Using these characteristics, the first stage was labeled as “Healthy.” Instances of this Healthy stage were observed in images obtained from both the ATP-induced eyes and the control eyes, as ATP-induced retinal degeneration is known to be patchy and nonuniform across the retina.[34–36]

The second stage was characterized to contain more disrupted features. The OS lost its straight extension feature from the ONL, in comparison to the Healthy stage. We note that the outer part of the photoreceptors points to the nodal point of the eye, and as such, their orientation will vary due to the section’s orientation but also the retinal location. The direction these features extended to was not considered an essential part of grading as this was dependent on the plane of section and retinal location. The ONL also showed signs of damage, evident as a decrease in thickness and/or density of nuclei. The lamination of the retinal layers was still well preserved and the RPE showed slight disruption such as cell migration to OS and ONL, all of which constituted the second stage named “Mild Damage” (Fig. 2b).

A third degeneration stage was defined when nearly all of the inner and outer photoreceptor segments were absent. The thickness of the ONL was considerably

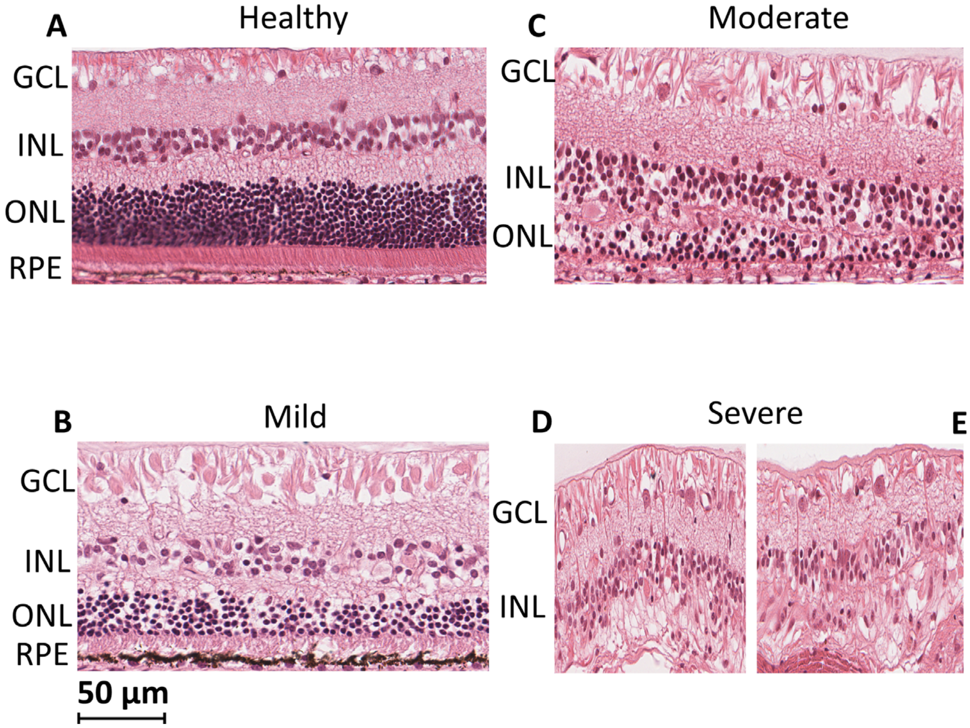

Figure 2. Examples of each degeneration stage, A) healthy, B) mild damage, C) moderate damage, D)–E) severe damage. The following annotations represent the retinal cellular layers in the segment; GCL: ganglion cell layer, INL: inner nuclear layer, ONL: outer nuclear layer, RPE: retinal pigment epithelium.

reduced compared to normal healthy tissue and the lamination of the layers was less distinct, for example, with nuclei present in the plexiform layers. This stage was labeled as “Moderate Damage” (Fig. 2c). The final stage of degeneration was characterized as a complete loss of the ONL, which led to an absence of clear lamination of the cellular layers. Mass cell migration and cell death were also visible in many cases in this stage. This last stage, labeled “Severe Damage,” included retinal areas where the INL and IPL were still observed (Fig. 2d) to retinal areas where the layering of the retina was unrecognizable (Fig. 2e).

Data Classification

Initially, a sample of 20 cropped images (5 from control eyes and 15 from ATP eyes) were given to two experts; one in retinal networks (MK) and another in retinal pathology (RW). The experts were asked to identify best sample images for the four different classes based on the criteria. This was considered an important step as we intended that all the remaining images for training and testing the AI model would be classified based on the same criteria and using the *ideal* example images as a guide. The 454 cropped images were split into two different sets; training and testing, which constituted 81% and 19% (369 and 85 images) of the preprocessed dataset, respectively.

Once *ideal* examples were determined, the training set images were then independently classified by two trained graders (DA and MS) into the four degeneration classes. Both graders were given the *ideal* example images to use as a guide and consulted the written criteria during classification. A retinal expert would adjudicate if the graders did not reach the same classification for any given image. Out of the 369 training set images, 67 required expert adjudication. The testing set was provided to six different observers; two retinal experts, the two trained graders who classified the training set, and two ophthalmology researchers (Table 2). Each observer classified the testing images independently and were asked to base their decision for each image on the *ideal* examples set and degeneration criteria.

Deep Learning Approach

A custom convolutional neural network (CNN) was developed and trained for the task of classifying the images into the four degeneration classes (Fig. 3). The model takes a preprocessed retinal image as an input and performs a series of mathematical operations

Table 1. The Degeneration Criteria Used to Implement the 4-Class Classification System

Stage Number	Stage Name	Biological Features
1	*Healthy*	Typical retinal layers observed (RPE, OS, ONL, OPL, INL, IPL, GCL), normal retinal lamination evident and the outer photoreceptors were organized. No sign of retinal damage (i.e., absence of a layer, low cell body density in nuclear layers, cell migration in plexiform layers). Some histological artifacts may be present but minimal. Some sections may be cut slightly obliquely.
2	*Mild damage*	Reduction in ONL thickness (compared to normal ONL thickness of 8–15 rows of nuclei) and/or nuclei density. Outer photoreceptor (outer and inner segment) and RPE disorganization as the previous class. Retinal lamination still observed with distinction between nuclear and plexiform layers evident.
3	*Moderate damage*	Large reduction in ONL thickness and/or nuclear density (over half that of the normal retina). Further degenerative alterations to the outer segments and RPE. Retinal lamination no longer preserved; signs of discontinuity between nuclear and plexiform layers.
4	*Severe damage*	Complete loss of outer retinal layers including ONL, outer segments and RPE. No clear lamination of nuclear and plexiform layers. Evidence of cell migration.

Table 2. Observer Description

Observer Number	Observer Alias	Observer Description
1	DA	Trained grader
2	MS	Trained grader and ophthalmology researcher
3	DN	Ophthalmology researcher
4	RW	Retinal pathology expert
5	LN-S	Ophthalmology researcher
6	MK	Retinal networks expert

on the image's pixels to finally predict its degeneration class. The model is made up of several interconnected convolutional layers, where two successive layers constitute a "block," a flattening layer that converts the convoluted image sequence into a vector that can be passed into two fully connected layers, before producing the maximum class probability indicating the model's prediction (Fig. 3).

A variation of hold-out was implemented to train the CNN on the training set, from which a 20% subset, namely a validation set, was chosen randomly for each training epoch. The following hyperparameters were maintained for all final experimentations: 2000 epochs for training, batch size of 128, learning rate of 1e-3, and a stochastic gradient descent optimizer. The number of convolutional blocks and the size of the filters in each layer were tested in a series of architecture search (AS) experiments (Table 3). Our AS tested different variations of the two architectural parameters, while keeping the remaining hyperparameters fixed. A total of 12 different architectures were trained and then evaluated on the unobserved testing set. Interobserver variability was also evaluated to compare with the best model's performance on the six observer testing sets. The code can be found online on GitHub (see Supplementary code).

Model Evaluation and Data Analysis

The model's training and testing performance was evaluated using a weighted-F1 score used to measure the accuracy of our model by compensating for the

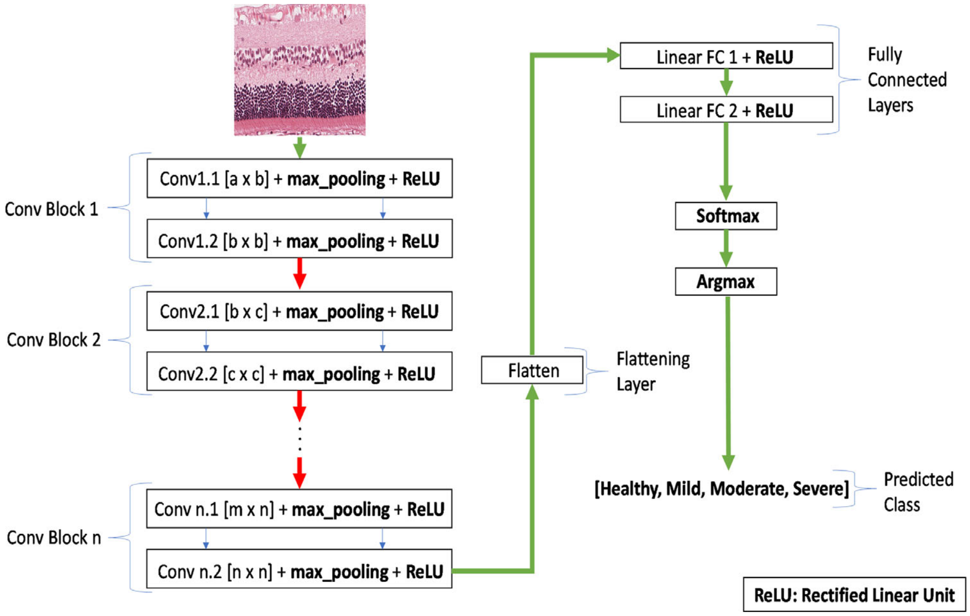
Conv Block 1
Conv1.1 [a x b] + max_pooling + ReLU
Conv1.2 [b x b] + max_pooling + ReLU
Conv Block 2
Conv2.1 [b x c] + max_pooling + ReLU
Conv2.2 [c x c] + max_pooling + ReLU
Conv Block n
Conv n.1 [m x n] + max_pooling + ReLU
Conv n.2 [n x n] + max_pooling + ReLU
Flatten
Flattening Layer
Linear FC 1 + ReLU
Linear FC 2 + ReLU
Fully Connected Layers
Softmax
Argmax
[Healthy, Mild, Moderate, Severe]
Predicted Class
ReLU: Rectified Linear Unit

Figure 3. A general schematic of the convolutional neural network's architecture and the classification's end-to-end workflow. Each layer's input was equal to the previous layer's output, while its output size was equal to its assigned block's convolutional filter size (Table 3).

Table 3. The 12 Different Architectures Used in the Architecture Search Experiments.

Architecture Label	Number of Convolutional Blocks	Output Filter Size of Each Convolutional Block
Arch_01	2	[64, 128]
Arch_02	3	[64, 96, 128]
Arch_03	4	[64, 96, 128, 256]
Arch_04	2	[128, 256]
Arch_05	3	[128, 256, 320]
Arch_06	4	[128, 256, 320, 512]
Arch_07	2	[256, 512]
Arch_08	3	[256, 320, 512]
Arch_09	4	[256, 320, 512, 1024]
Arch_10	2	[512, 1024]
Arch_11	3	[512, 720, 1024]
Arch_12	4	[256, 512, 720, 1024]

imbalanced distribution of class data instances, while evaluating the balance between precision and recall (Eq. 1). To calculate an F1 score, either the observer classification or the model's prediction has to be designated as the true label. A total of five training rounds per configuration were run and a mean weighted-F1 score was calculated. The model with the highest mean weighted-F1 score across all observers was selected as the best model. True accuracy, defined as the number of true predictions (true positives and true negatives) out of all predictions, and cross entropy loss (log loss), a metric used to evaluate the model's divergence from the actual true classification, as a function of the number of epochs were then visualized for the best model. For the training set F1 score, only the training (and validation) grader labels were defined as the true labels for the classification, hence requiring one weighted-F1 score to be measured. For the six observer testing sets, two weighted-F1 scores were calculated, once when the model's prediction labels were designated to be the true labels, and another when the observer's labels were designated as such.

$$weighted\ F1 = \frac{\sum_i^{classes} \left(F_{n_i} + T_{p_i}\right) .(F1_i)}{N} \tag{1}$$

where:

$$F1_i = \frac{2\,(precision_i)\,(recall_i)}{precision_i + recall_i},$$

$$precision_i = \frac{T_{p_i}}{T_{p_i} + F_{p_i}},$$

$$recall_i = \frac{T_{p_i}}{T_{p_i} + F_{n_i}}$$

T_{p_i}: True positive instances of class i, F_{p_i}: False positive instances of class i, F_{n_i}: False negative instances of class i, N: Total number of samples

Equation 1: Weighted-F1 equations

Interobserver variability was visualized using confusion matrices, which capture the raw agreement between any two observers (including the model). The variability was quantified using Cohen's kappa coefficient,[37] as it is standardized to produce a score in the range [0–1] and can directly evaluate interobserver agreement (Eq. 2). A substantially high level of agreement between the observers can be inferred from kappa scores in the range [0.8–1] and lower scores of [0.6–0.79] and [0–0.59] indicate a moderate and low degree of agreement, respectively. Weighted-F1 scores between the observers were also measured to reinforce the kappa measurements and further validate the degree of observer agreement. Heatmaps of the testing set image classifications were generated to demonstrate the raw predictions made by all observers and were compared to the model's prediction.

$$\kappa = \frac{Pr\,(a) - Pr\,(e)}{1 - Pr\,(e)} \tag{2}$$

where:

$$Pr(a) = \frac{\sum_i^{Classes} \sum T_{p_i}}{n},$$

$$Pr\,(e) = \frac{\sum_i^{Classes} \left(\frac{(F_{P_i} + T_{P_i})(F_{n_i} + T_{P_i})}{n} \right)}{n}$$

T_{p_i}: True positive instances of class *i*, F_{p_i}: False positive instances of class *i*, F_{n_i}: False negative instances of class *i*

Equation 2: Cohen's Kappa Score equations

Results

Architecture Search

Of the 12 different architectures tested, all architectures performed consistently well and Arch_04 was found to have the highest mean weighted-F1 score of 84.4% across all observers (Fig. 4). This configuration was selected as the best model for subsequent analyses and was renamed RDP-Net-Ret (Retinal Degeneration Prediction Network—Retina). For simplicity, this best model will be referred to as RDP-Net in this study. Its training and validation true accuracies converged to 1 and 0.928, respectively (Fig. 5a), and losses converged to 0.04 and 0.2, respectively (Fig. 5b). It is also interesting to note that the second and third best architectures were Arch_05 and Arch_01, both of which were among the less complex architectures considered.

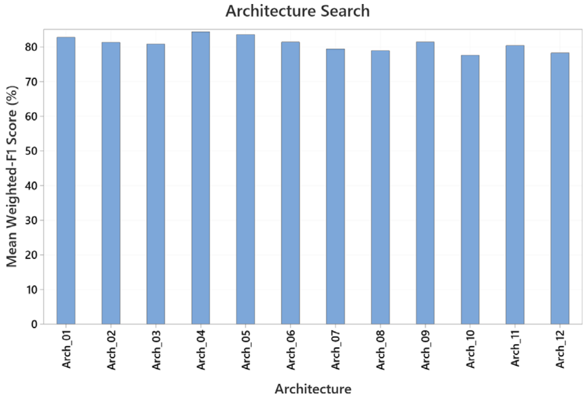

Figure 4. The mean weighted-F1 scores on the testing set for all architectures across the six observers.

Interobserver Variability

Confusion matrices between RPD-Net and two of the six observers (DA and MS) who were also the training set graders, demonstrated high agreement (high diagonal values in the matrices) regardless of whether RDP-Net's grading or the observers' grading were designated as the true labels and across all four classes of retinal degeneration. Confusion matrices for all other observers were similar (Supplementary Figures S1 and S2) and there were no systematic confusions between the observers and RDP-Net.

When viewing the raw agreement between the observers and the model, we found that the model as well as nearly all the observers labeled most images the same but interobserver variability was clearly present, especially for images showing some level of damage (Fig. 7a). Observer 5 was an exception, who tended to be more conservative when classifying images into the three damage classes, as several images were considered by this observer to be that of healthy retina (Fig. 7a), but they were classified as mild damage by other observers and the model. Another interesting observation is that all predicted misclassifications by the model (except for one image) were consistently one class above or below the observer's true label (Fig. 6, Fig. 7a). The

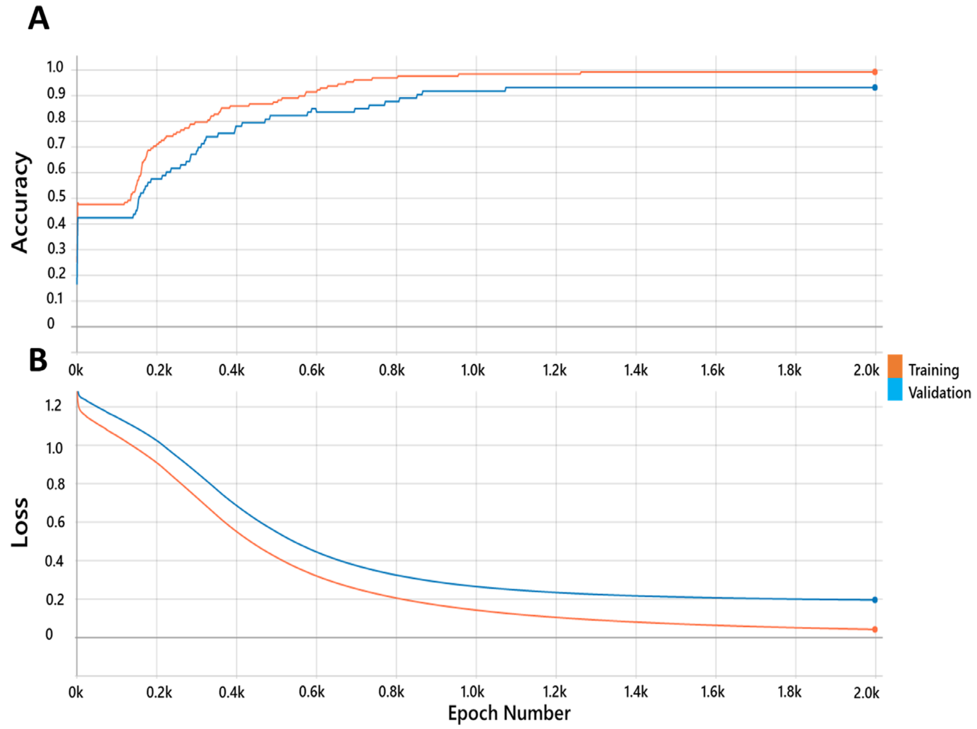

Figure 5. Training and validation A) true accuracy and B) cross entropy log loss curve for RDP-Net.

overall agreements between the model and observers, and between pairs of observers were quantified with the kappa score (Fig. 7b). When comparing RDP-Net's level of agreement with other observers, kappa scores varied between 0.59–0.86 (M = 0.7633, SD = 0.0961) while the interobserver kappa scores varied between 0.56–0.91 (M = 0.771, SD = 0.109). Both groups of kappa scores were not significantly different from each other (*t*-test, $t(10) = -0.15$, $P = 0.88$) (Fig. 7b). Similarly, the weighted-F1 scores between RDP-Net and the individual observers varied between 76% to 91%, and 70% to 94%, depending on which values were taken as the true labels, once again demonstrating equivalence between model-observer variability and interobserver variability (Fig. 7c).

Data Experimentation

To observe changes in RDP-Net's performance as the input data were changed, we undertook a series of experiments in which several data parameters were varied, and the model performance was re-evaluated. The first of these experiments involved retraining the model on the same training images when they were halved to 125 μm in width, with the number of these images doubling as a result, to evaluate RDP-Net's robustness when dealing with less retinal context. Once retrained, the model agreement against the original observer testing set classifications was recalculated. RDP-Net showed no significant difference in kappa scores when trained on the original training set or the halved training set ($t(15) = -1.3$, $P = 0.21$) (Fig. 7b vs Fig. 8a). A heatmap and weighted-F1 scores for this experiment (similar to Figs. 7a and 7c) displayed similar results to the overall kappa scores (Supplementary Fig. S3).

The second experiment involved keeping the cropped images for training as full width but reducing the training set size by randomly eliminating images from the training set, to determine how much training played a role in RDP-Net's performance. This reduction was done in 10% steps down to a reduction of 30% (i.e., the model was trained on sizes of 90%, 80%, and 70% of the original training set). The reduction was limited to 30% as beyond this value, further reduction would require modification of certain hyperparameters, which would change the model architecture. Once retrained, the model agreement against the original observer classifications was recalculated. A gradual decline in agreement performance was observed across all observers as the training set size was reduced. A one-way ANOVA comparing the mean performance across the six observers for the four training set sizes, showed that performance was significantly different

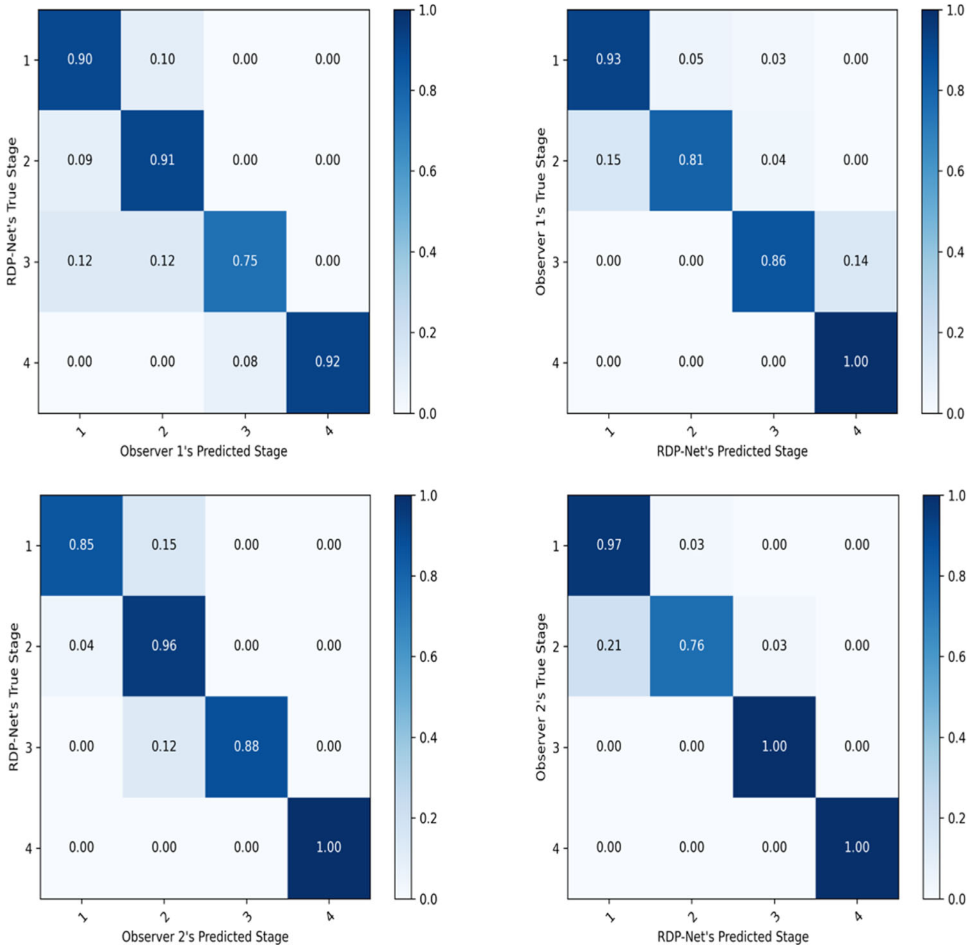

Figure 6. The confusion matrices which measured the agreement between; A) Observer 1's (DA) predictions and RDP-Net's labels, B) Observer 2's (MS) predictions and RDP-Net's labels, C) RDP-Net's predictions and Observer 1's true labels, D) RDP-Net's predictions and Observer 2's true labels.

for at least one pair of different training set sizes ($F_{3,20} = 7.28$, $P = 0.002$, $\eta_p{}^2 = 0.52$). Tukey post hoc tests revealed that there were two training size pairs (100%, 70%) and (90%, 70%) which resulted in no significant performance difference (Fig. 8b), while other pairs showed no significance.

Discussion

The present study showed that a convolutional neural network can accurately and reliably classify histologically stained retinal images into separate predefined degeneration classes. This, to our knowledge, is the first attempt to use AI for such a task. RDP-Net's training and validation metrics indicated that it reached its maximum potential using the chosen architectural parameters, with no signs of under- or overfitting. RDP-Net was highly accurate with no significant differences in interobserver versus RDP-Net and observers' kappa scores. RDP-Net also demonstrated robustness with similar levels of agreement to observers when trained on images with less context. However, these results were not replicated when the number of training images was reduced, suggesting RDP-Net's performance is more dependent on the quantity, rather than the context of the training set.

Model Complexity Versus Performance

Certain architectures, like CNNs, are known to be effective in image classification tasks, as their highly tunable complexity offers unprecedented access for machines to learn without the hassle of *feature engineering*.[38] Although this is highly advantageous for medical imaging tasks, which require vast expert

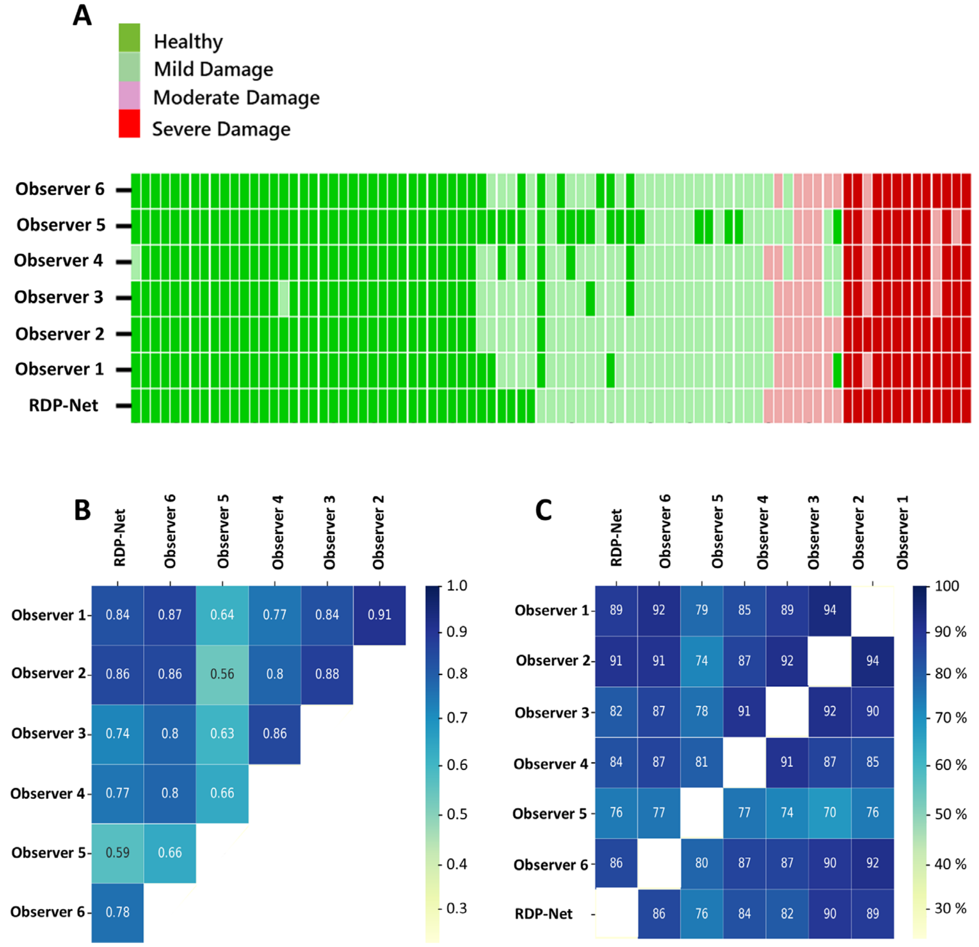

Figure 7. Interobserver variability plots: A) The agreement heatmap between all observers and RDP-Net for the 85 testing set images, where each column represents a single image and a row represents an observer's classification for all 85 images. B) The Cohen kappa scores between all observers and RDP-Net. C) The weighted-F1 scores between the observers and RDP-Net, with the columns representing the *true* labels.

resources to process the relevant features, its current multitude indicates that the search space for an optimal model capable of performing expert-level classification is exponentially large.[39] Additionally, many architectural parameters, which control how a neural network may perform, further exacerbate this search space. Our three best performing configurations (Arch_04, Arch_05, and Arch_01) showed that (1) the top performing models were not the most complex architectures tested, (2) the best performing model, Arch_04 or RDP-Net, was the second simplest architecture out of the top three performing models and (3) the search space's local maximum, which was derived from the 12 tested architectures, could potentially be a global maximum for this specific classification task (Fig. 4).

When compared to more prominent and well-known architectures such as ResNET and Inception used in image classification with nonretinal histological[40–42] and nonhistological images,[15,43,44] RDP-Net was far less complex (two layers as opposed to 18–27 layer architectures), hinting at the notion that not only are the visual features within histological data easily categorized by an algorithm, but the well-segmented structure of the retina and its easily distinguishable features when stained with H&E, could have played a large role in reducing the complexity of RDP-Net. This prompts the need for further investigation into deep learning parameter optimization for models such as RDP-Net coupled with similar data to prove these hypotheses.

Model and Observer Classifications Reflect a Continuum of Retinal Degeneration

Another interesting observation from our results is that all predicted misclassifications by the model (except for one image) were consistently one class above

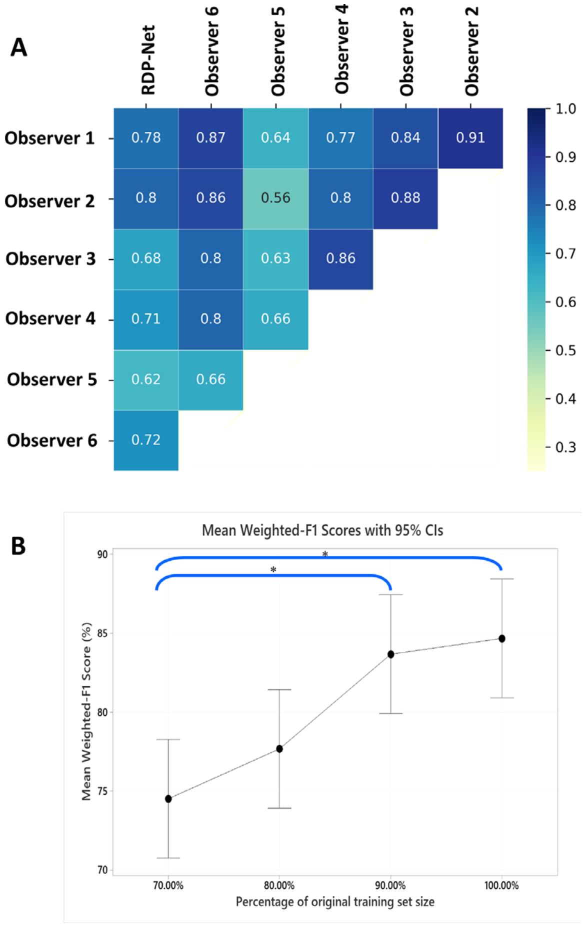

Figure 8. (A) The Cohen kappa scores between all observers and RDP-Net when trained on a new training set with half the width of the original images. B) The mean weighted-F1 scores when RDP-Net was trained on different sizes of the original training set. An asterisk (*) indicates significant difference between the two associated groups (Tukey test, $P < 0.05$). The error bars represent standard errors.

or below the observer's true label (Fig. 6, Fig. 7b). In addition, certain images for which the observers classified them differently, when asked about their classification attempt, responded that these images contained one or more features from two adjacent classes but they in the end had to choose one class. The reasons for these apparent nonagreements suggest that retinal degeneration stages are not necessarily discrete. One can think of these *contentious* images to be representatives of *degenerative gradients* in which the retina may be transitioning into a higher degenerative stage and therefore could be considered as belonging to *hidden* stages that exist between the present defined stages. It is possible that further degeneration stages could be defined to fine grain the boundaries between the existing stages, and guide future studies to solidify universally accepted degeneration stages, but ultimately there may always be a continuum of degeneration rather than precisely discrete and independent stages.[30] Another possible way to get around this issue is by tweaking the surrounding context of input and thus directly influencing training of the model, resulting in a more well-defined and sensible representation of the different retinal degeneration stages. Indeed, reducing the number of neighboring features in a single data instance does not significantly affect the model's performance, and thus, this can then be extended to an optimization problem where the search space of these different data parameters can be fine-tuned collectively and produce an even more powerful automated tool.

Further Applicability of RDP-Net

Our results demonstrate that with the current architecture of RDP-Net, a scenario in which a robust *digital copy* of an observer can be developed and implemented in practice is highly probable since the model performed best when compared against the two observers responsible for grading the training set. Future iterations of the model could be generalized enough so it conforms to another observer's expectations, regardless of that observer's involvement in the *degeneration criteria* design and training set classification. This would not only result in significant time and resource savings when analyzing large datasets but could also aid in standardization of criteria used to assess retinal degeneration across individual researchers.

Secondly, while the model in this study was applied to H&E images, the current version of the model could be adapted to other histological data or adapted to quantify features such as cell density, retinal thickness, or degree of antibody staining. Its portability will need to be tested before applying it on other histological data, but it certainly has the potential to be used as a transfer learning model. Other data may illustrate different visual features more prominently, in which case it may require the model to undergo a new architecture search experiment to evaluate the performance of different configurations for the set task, using the appropriate training and testing data.

Thirdly, it is possible, and indeed likely, that the personal and professional interactions between pairs of observers over more than a decade of collaboration influenced the manner in which degeneration criteria were weighted, further influencing their classifications. This is perhaps best exemplified by the relatively lower interobserver variability between observers 3 and 4 (who have worked together for over a decade), as well as 3 and 1, and 1 and 2 compared to other pairs of observers. This suggests that a tacit diffusion of shared understanding may underpin the transferal of information with the human network and future versions of the model could take advantage of this phenomenon by having the model trained on degeneration criteria that are standardized across several like-minded experts.

Model Interpretability

While RDP-Net demonstrates that it is capable of performing this classification task with reasonably high accuracy comparable to that of the observers, it is important to understand the level of explainability that can be extracted from its performance. This is currently an active area of research, due to the notorious black box nature of CNNs.[45,46] Despite this, our drafted criteria and *ideal* example images serve as a stepping stone into our model's interpretability, as its accuracy is preserved when attempting to classify different classes of degeneration. The feature space on which RDP-Net makes its predictions intersects largely, if not completely, with the proposed criteria features, as our results demonstrate. This is important as previous attempts to incorporate AI into medicine lacked this level of transparency when certain models would completely misinterpret data.[47] However, the experiments to exactly determine what these features are, are out of this study's scope. Additionally, mechanically quantifying these particular visual features, for example, the nuclear density in different retinal layers or the thickness of these layers, may further solidify the exact feature space on which the model is searching. Rather than experts giving subjective opinions when evaluating these visual features, fully quantifiable and more informative metrics may assist them, and by extension the model, to make more informed decisions about the retina's cellular condition. Hence, identifying precisely, and perhaps quantifying, such feature spaces will be highly important, as clinicians and researchers will be able to establish new ways to incorporate AI into pipelines safely with high confidence.

Limitations

This study only used one animal model of retinal degeneration to train and test RDP-Net and therefore the performance of RDP-Net for other retinal degenerations is not clear. In particular, our dataset was derived from the feline retina and so species-specific differences in retinal anatomy could affect its performance for other datasets. A large amount of evidence, however, suggests anatomical changes such as cell migration, synaptic remodeling, and cell loss, that are most likely to have been used as features by this model to perform classifications, are common to other models of photoreceptor degeneration, including in humans.[26] In addition, the inherent patchy nature of degeneration present in our chosen degeneration model provided the advantage of being able to encounter multiple degeneration stages within the same retinal sample. Future studies, however, should utilize datasets from different models of retinal degeneration preferably with several different etiologies to confirm the accuracy and robustness of the model. Computationally, the experimental search space could also be increased to explore more architectures and increase the accuracy of our model.

Conclusions

This study demonstrates proof of concept of a novel attempt to classify retinal degeneration on a cellular level using AI. Our results indicate that the development of *digital experts*, that can be personalized to specific experts, while maintaining a certain degree of generalizability across other expert opinions, is realistic. Our study lays the foundation for implementing powerful automated frameworks that can alleviate resource heavy tasks such as interpretation of histological data and provide meaningful and objective insights into retinal disease.

Supplementary Data Code

Code used for architecture training/testing can be found here: https://github.com/dalmouiee/RDP-Net-Ret.

Raw data (image training and testing datasets) can be made available on request.

Acknowledgments

The authors thank the staff of the Biological Resource Centre, University of Melbourne at the Royal Victorian Eye and Ear Hospital for assistance with animal husbandry. Funding for conducting the animal experiments was provided by the National Health and Medical Research Council (NHMRC) Project Grant (GNT1063093). The Bionics Institute acknowledges the support it receives from the Victorian Government through its Operational Infrastructure Support Program.

Diclosure: **D. Al Mouiee**, None; **E. Meijering**, None; **M. Kalloniatis**, None; **L. Nivison-Smith**, None; **R.A. Williams**, None; **D.A.X. Nayagam**, None; **T.C. Spencer**, None; **C.D. Luu**, None; **C. McGowan**, None; **S.B. Epp**, None; **M.N. Shivdasani**, None

References

1. Jones B, Kondo M, Terasaki H, Lin Y, McCall M, Marc R. Retinal remodeling. *Jpn J Ophthalmol*. 2012;56(4):289–306.
2. Jones B, Pfeiffer R, Ferrell W, Watt C, Marmor M, Marc R. Retinal remodeling in human retinitis pigmentosa. *Exp Eye Res*. 2016;150:149–165.
3. Jones BW, Marc RE. Retinal remodeling during retinal degeneration. *Exp Eye Res*. 2005;81(2):123–137.
4. Jones BW, Marc RE, Pfeiffer RL. Retinal degeneration, remodeling and plasticity. In: Kolb H, Fernandez E, Nelson R, eds. *Webvision: The Organization of the Retina and Visual System [Internet]*. Moran Eye Center: University of Utah Health Sciences Center; 2016.
5. Marc RE, Jones BW, Anderson JR, et al. Neural reprogramming in retinal degeneration. *Invest Ophthalmol Vis Sci*. 2007;48(7):3364–3371.
6. Marc RE, Jones BW, Watt CB, Strettoi E. Neural remodeling in retinal degeneration. *Prog Retin Eye Res*. 2003;22(5):607–655.
7. Pfeiffer RL, Marc RE, Jones BW. Persistent remodeling and neurodegeneration in late-stage retinal degeneration. *Prog Retin Eye Res*. 2020;74:100771.
8. Chua J, Fletcher EL, Kalloniatis M. Functional remodeling of glutamate receptors by inner retinal neurons occurs from an early stage of retinal degeneration. *J Comp Neurol*. 2009;514(5):473–491.
9. Chua J, Nivison-Smith L, Fletcher EL, Trenholm S, Awatramani GB, Kalloniatis M. Early remodeling of Muller cells in the rd/rd mouse model of retinal dystrophy. *J Comp Neurol*. 2013;521(11):2439–2453.
10. Marc R, Pfeiffer R, Jones B. Retinal prosthetics, optogenetics, and chemical photoswitches. *ACS Chem Neurosci*. 2014;5(10):895–901.
11. Lim LL, Suhler EB, Rosenbaum JT, Wilson DJ. The role of choroidal and retinal biopsies in the diagnosis and management of atypical presentations of uveitis. *Trans Am Ophthalmol Soc*. 2005;103:84.
12. Rishi P, Dhami A, Biswas J. Biopsy techniques for intraocular tumors. *Indian J Ophthalmol*. 2016;64(6):415.
13. Hiscott P, Wong D, Grierson I. Challenges in ophthalmic pathology: the vitreoretinal membrane biopsy. *Eye*. 2000;14(4):549–559.
14. Westerfeld C, Mukai S. Retinal and choroidal biopsy. *Int Ophthalmol Clin*. 2009;49(1):145–154.
15. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402–2410.
16. Levenkova A, Sowmya A, Kalloniatis M, Ly A, Ho A. Automatic detection of diabetic retinopathy features in ultra-wide field retinal images. *Proc. SPIE 10134, Medical Imaging 2017: Computer-Aided Diagnosis, 101341M*. 2017, https://doi.org/10.1117/12.2253980.
17. Kapoor R, Walters SP, Al-Aswad LA. The current state of artificial intelligence in ophthalmology. *Surv Ophthalmol*. 2019;64(2):233–240.
18. Kapoor R, Whigham BT, Al-Aswad LA. Artificial intelligence and optical coherence tomography imaging. *Asia Pac J Ophthalmol*. 2019;8(2):187–194.
19. Russakoff DB, Lamin A, Oakley JD, Dubis AM, Sivaprasad S. Deep learning for prediction of AMD progression: a pilot study. *Invest Ophthalmol Vis Sci*. 2019;60(2):712–722.
20. Lu W, Tong Y, Yu Y, Xing Y, Chen C, Shen Y. Deep learning-based automated classification of multi-categorical abnormalities from optical coherence tomography images. *Transl Vis Sci Technol*. 2018;7(6):41.
21. Djuric U, Zadeh G, Aldape K, Diamandis P. Precision histology: how deep learning is poised to revitalize histomorphology for personalized cancer care. *NPJ Precis Oncol*. 2017;1(1):1–5.
22. Hosseini S, Chen H, Jablonski MM. Automatic detection and counting of retina cell nuclei using deep learning. *arXiv*. preprint arXiv:200203563. 2020.
23. Liu Y, Kohlberger T, Norouzi M,, *et al.* Artificial intelligence–based breast cancer nodal metastasis detection: insights into the black box for pathologists. *Arch Pathol Lab Med*. 2019;143(7):859–868.
24. Lui TK, Wong KK, Mak LL, et al. Feedback from artificial intelligence improved the learning of junior endoscopists on histology prediction of gastric lesions. *Endosc Int Open*. 2020;8(2):E139–E146.

25. Helmstaedter M, Briggman KL, Turaga SC, Jain V, Seung HS, Denk W. Connectomic reconstruction of the inner plexiform layer in the mouse retina. *Nature*. 2013;500(7461):168–174.
26. Kalloniatis M, Nivison-Smith L, Chua J, Acosta M, Fletcher E. Using the rd1 mouse to understand functional and anatomical retinal remodelling and treatment implications in retinitis pigmentosa: a review. *Exp Eye Res*. 2016;150:106–121.
27. Halupka KJ, Abbott CJ, Wong YT, et al. Neural responses to multielectrode stimulation of healthy and degenerate retina. *Invest Ophthalmol Vis Sci*. 2017;58(9):3770–3784.
28. Spencer TC, Fallon JB, Abbott CJ, Allen PJ, Brandli A, Luu CD, Epp SB, Shivdasani MN. Electrical field shaping techniques in a feline model of retinal degeneration. *Annu Int Conf IEEE Eng Med Biol Soc*. 2018:1222–1225, doi:10.1109/EMBC.2018.8512473.
29. Nayagam DA, McGowan C, Villalobos J, et al. Techniques for processing eyes implanted with a retinal prosthesis for localized histopathological analysis. *J Vis Exp*. 2013(78):e50411.
30. Leung RT, Shivdasani MN, Nayagam DAX, Shepherd RK. In vivo and in vitro comparison of the charge injection capacity of platinum macroelectrodes. *IEEE Trans Biomed Eng*. 2015;62(3):849–857.
31. Villalobos J, Fallon JB, Nayagam DA, et al. Cortical activation following chronic passive implantation of a wide-field suprachoroidal retinal prosthesis. *J Neural Eng*. 2014;11(4):046017.
32. Bankhead P, Loughrey MB, Fernández JA, et al. QuPath: open source software for digital pathology image analysis. *Sci Rep*. 2017;7(1):1–7.
33. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. *Nat Methods*. 2012;9(7):671–675.
34. Aplin FP, Fletcher EL, Luu CD, et al. Stimulation of a suprachoroidal retinal prosthesis drives cortical responses in a feline model of retinal degeneration. *Invest Ophthalmol Vis Sci*. 2016;57(13):5216–2529.
35. Aplin FP, Luu CD, Vessey KA, Guymer RH, Shepherd RK, Fletcher EL. ATP-induced photoreceptor death in a feline model of retinal degeneration. *Invest Ophthalmol Vis Sci*. 2014;55(12):8319–8329.
36. Aplin FP, Vessey KA, Luu CD, Guymer RH, Shepherd RK, Fletcher EL. Retinal changes in an ATP-induced model of retinal degeneration. *Front Neuroanat*. 2016;10:46.
37. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*. 2012;22(3):276–282.
38. Jiang Y, et al. Expert feature-engineering vs. deep neural networks: Which is better for sensor-free affect detection?. In: Penstein Rosé C, et al. eds. *Artificial Intelligence in Education. AIED 2018. Lecture Notes in Computer Science, vol 10947*. Springer, Cham. 2018, https://doi.org/10.1007/978-3-319-93843-1_15.
39. Perone CS, Cohen-Adad J. Promises and limitations of deep learning for medical image segmentation. *J Med Artif Intell*. 2019;2:1, doi:10.21037/jmai.2019.01.0.
40. Iizuka O, Kanavati F, Kato K, Rambeau M, Arihiro K, Tsuneki M. Deep Learning models for histopathological classification of gastric and colonic epithelial tumours. *Sci Rep*. 2020;10(1):1–11.
41. Kather JN, Pearson AT, Halama N, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med*. 2019;25(7):1054–1056.
42. Roy K, Banik D, Bhattacharjee D, Nasipuri M. Patch-based system for classification of breast histology images using deep learning. *Comput Med Imaging Graph*. 2019;71:90–103.
43. Guo S, Yang Z. Multi-channel-ResNet: an integration framework towards skin lesion analysis. *Inform Med Unlocked*. 2018;12:67–74.
44. Zhang Z, Wu C, Coleman S, Kerr D. DENSE-INception U-net for medical image segmentation. *Comput Methods Programs Biomed*. 2020;192:105395.
45. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, LJae-p Kagal. Explaining explanations: an overview of interpretability of machine learning, 2018: [arXiv:1806.00069 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2018arXiv180600069G.
46. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev Data Min Knowl Discov*. 2019;9(4):e1312.
47. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*. 2018;6:52138–52160.