OPEN ACCESS

Clinical science

Ensemble of deep convolutional neural networks is more accurate and reliable than board-certified ophthalmologists at detecting multiple diseases in retinal fundus photographs

Prashant U Pandey ,[1] Brian G Ballios,[2,3,4] Panos G Christakis,[2,4] Alexander J Kaplan,[2] David J Mathew ,[2,3,4] Stephan Ong Tone,[2,5] Michael J Wan ,[2] Jonathan A Micieli,[2,4,6] Jovi C Y Wong[2]

[1]School of Biomedical Engineering, The University of British Columbia, Vancouver, British Columbia, Canada
[2]Department of Ophthalmology and Vision Sciences, University of Toronto, Toronto, Ontario, Canada
[3]Krembil Research Institute, University Health Network, Toronto, Ontario, Canada
[4]Kensington Vision and Research Centre and Kensington Research Institute, Toronto, Ontario, Canada
[5]Sunnybrook Research Institute, Toronto, Ontario, Canada
[6]Department of Ophthalmology, St. Michael's Hospital, Unity Health, Toronto, Ontario, Canada

Correspondence to
Dr Jovi C Y Wong, Department of Ophthalmology and Vision Sciences, University of Toronto, Toronto, Canada; jovi.wong@mail.utoronto.ca

Received 7 July 2022
Accepted 11 January 2023
Published Online First 31 January 2023

To cite: Pandey PU, Ballios BG, Christakis PG, *et al*. *Br J Ophthalmol* 2024;**108**:417–423.

ABSTRACT

Aims To develop an algorithm to classify multiple retinal pathologies accurately and reliably from fundus photographs and to validate its performance against human experts.

Methods We trained a deep convolutional ensemble (DCE), an ensemble of five convolutional neural networks (CNNs), to classify retinal fundus photographs into diabetic retinopathy (DR), glaucoma, age-related macular degeneration (AMD) and normal eyes. The CNN architecture was based on the InceptionV3 model, and initial weights were pretrained on the ImageNet dataset. We used 43 055 fundus images from 12 public datasets. Five trained ensembles were then tested on an 'unseen' set of 100 images. Seven board-certified ophthalmologists were asked to classify these test images.

Results Board-certified ophthalmologists achieved a mean accuracy of 72.7% over all classes, while the DCE achieved a mean accuracy of 79.2% (p=0.03). The DCE had a statistically significant higher mean F1-score for DR classification compared with the ophthalmologists (76.8% vs 57.5%; p=0.01) and greater but statistically non-significant mean F1-scores for glaucoma (83.9% vs 75.7%; p=0.10), AMD (85.9% vs 85.2%; p=0.69) and normal eyes (73.0% vs 70.5%; p=0.39). The DCE had a greater mean agreement between accuracy and confident of 81.6% vs 70.3% (p<0.001).

Discussion We developed a deep learning model and found that it could more accurately and reliably classify four categories of fundus images compared with board-certified ophthalmologists. This work provides proof-of-principle that an algorithm is capable of accurate and reliable recognition of multiple retinal diseases using only fundus photographs.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Artificial intelligence (AI) algorithms have demonstrated excellent accuracy in classifying pathologies from retinal fundus photographs.

WHAT THIS STUDY ADDS

⇒ Our AI algorithm demonstrates not only superior accuracy to board-certified ophthalmologists, in a balanced test containing four image categories, but also superior reliability as the confidence output by our model more closely matches its accuracy compared with ophthalmologists.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This model could be used a blueprint for future decision-making support systems to assist pathology detection both in specialist ophthalmology clinics and in generic healthcare settings such as family practices and emergency rooms.

INTRODUCTION

Retinal imaging plays a key role in the diagnosis of retinal pathologies. In current clinical practices, retinal imaging is manually interpreted by ophthalmologists and this workflow is limited by human resources. Automatic recognition of pathologies from fundus images would increase the efficiency in eye clinics, as well as introduce the potential of retinal screening in geographical regions where there is limited or infrequent access to specialists. In particular, several machine learning approaches based on convolutional neural networks (CNNs) have already been developed to recognise pathologies in fundus images.[1] Many of these methods are designed to classify only one category against normal samples, such as for diabetic retinopathy (DR) classification,[2] for papilloedema classification[3] and for glaucoma classification.[4] Recently, two such learning algorithms were granted clearance by the US Food and Drug Administration (FDA) for DR screening[5] and DR and diabetic macular oedema screening,[6] making them among the first diagnostic machine learning methods to be authorised by the FDA without the need for human oversight.

CNNs have also been used for grading pathologies on a nominal scale from fundus photographs, such as for age-related macular degeneration (AMD),[7] and recently studies have demonstrated that CNNs are capable of accurately detecting

multiple different retinal pathologies. Cen *et al* developed a two-level hierarchical classification technique for classifying 39 different categories of retinal conditions and found that it achieved comparable performance to five retinal specialists.[8] Similarly, Son *et al* trained 12 independent networks to detect 12 retinal findings in fundus images and found this technique performed equivalently to three retinal specialists in identifying haemorrhages and hard exudates.[9] Li *et al* developed an ensemble of CNNs to classify DR and diabetic macular oedema and demonstrated that it performed either as well or better than eight expert raters.[10] Ting *et al* trained and validated a CNN for detecting referable DR, possible glaucoma and AMD in approximately 500 000 retinal images from a large multiethnic population, achieving areas under the receiver operating characteristic curves (AUROCs) ranging from 0.931 to 0.983.[11] Detecting diseased retinal images using unsupervised anomaly detection was also proposed by Han *et al*, through the use of a convolutional generative adversarial network, which achieved an AUROC of 0.896 for detecting abnormal fundus images.[12] Zapata *et al* used five separate CNNs for five tasks such as differentiating optical coherence tomography (OCT) images from colour fundus photographs, classifying right eye (OD) from left eye (OS) images and detecting AMD and glaucomatous optic neuropathy.[13] The latter two networks achieved mean accuracies of 86.3% and 80.3%, respectively.

In order for stand-alone deployment, an automated retinal screening method should be able to identify multiple possible retinal pathologies. Moreover, the method should demonstrate equivalent or superior performance to the current standard-of-care in retinal diagnoses and should produce trustworthy predictions that can be used by clinicians. We developed a method and associated study to address these gaps with three contributions: (1) we trained an ensemble-based deep learning algorithm (deep convolutional ensemble: DCE) which is capable of detecting three major retinal pathologies and normal eyes from fundus images alone, (2) we directly compared the performance of the DCE against practising board-certified ophthalmologists on a balanced test set, to show it is overall more accurate and (3) we demonstrated that the output of the DCE is also more reliable than the ophthalmologists over the same set of images.

METHODS

Dataset

We compiled our training and validation image sets from 12 publicly available retinal fundus datasets[8 14–24] containing images of DR, glaucoma, AMD and normal patients (table 1). In total, the combined set contained 43 055 images, including 30 475 normal, 11 814 DR, 544 glaucoma and 222 AMD images. These images were separated into training and validation sets using an 80%/20% split. We created a separate test set by randomly sampling the aforementioned public datasets such that there were 25 images of each category, for a total of 100 test images. We ensured there was no more than one image per patient in the test set and that there was no overlap of patients and images between the training, validation and test sets. We used the disease classes directly as determined by each institution associated with the dataset and did not reclassify any images. For public datasets that originally included gradations of diseases, we combined any subclassifications into one overall disease class (ie, mild or severe DR was considered DR).

Deep convolutional ensemble

We implemented a DCE: a CNN-based ensemble classifier trained to predict the disease class in fundus images. The ensemble consisted of five InceptionV3[25] networks that were pretrained on the ImageNet dataset (figure 1). Each InceptionV3 model was independently trained on bootstrap aggregated samples from the training set, consistent with the deep ensembling methodology to improve uncertainty estimation and confidence calibration.[26 27] We trained using a weighted cross-entropy loss where the weights for each class were inversely proportional to the count of images in that class. We used the rectified Adam for optimisation and a fixed batch size of 68 images. Input images were resized to 299×299 pixels, and random horizontal flipping and random scaling between 0% and 10% were used for data augmentation during training. The final predicted class per image was generated by taking the majority vote of the five networks, such that the model could only predict one class per image. In the case there was no majority vote, we randomly assigned the predicted class from one of the categories with the most votes so as not to favour one class over the others. The network architecture and optimisation process were implemented in PyTorch and executed on a single Nvidia V100 GPU.

Table 1 Number of images in the training and validation set and in the test set, and the corresponding source datasets

	Training and validation set				Test set			
Source dataset	Normal	DR	Glaucoma	AMD	Normal	DR	Glaucoma	AMD
DiaretDB[14]	0	89	0	0	0	0	0	0
Drishti-GS[15]	31	0	68	0	0	0	2	0
DRIVE[16]	33	7	0	0	0	0	0	0
HRF[17]	15	15	13	0	0	0	2	0
IDRiD[18]	167	348	0	0	1	0	0	0
Kaggle-39[8]	38	105	0	0	0	1	0	0
Kaggle-DR[19]	25 769	9278	0	0	21	20	0	0
ODIR[20]	2276	1187	189	182	2	1	7	18
MESSIDOR[21]	546	651	0	0	0	3	0	0
ORIGA-light[22]	482	0	158	0	0	0	10	0
REFUGE[23]	1079	0	116	0	1	0	4	0
STARE[24]	39	134	0	40	0	0	0	7
Total	30 475	11 814	544	222	25	25	25	25

AMD, age-related macular degeneration; DR, diabetic retinopathy.

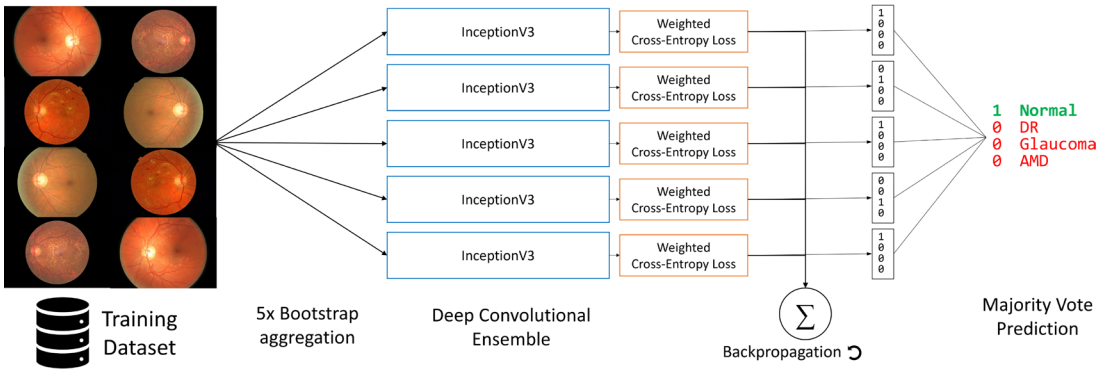

Figure 1 Overview of the deep convolutional ensemble model components and training. AMD, age-related macular degeneration; DR,diabetic retinopathy.

Confidence

To compare the reliability of the DCE to that of board-certified ophthalmologists, we also estimated the confidence of the ensembled models by taking the softmax of the mean logit output per image, and thresholding this value above 50% as 'confident' and below 50% as 'not confident'. This confidence estimation was not used during training.

Experiment on test data

Figure 2 illustrates the overall experiment process.

Deep convolutional ensemble

We trained the DCE for 20 epochs on the training set, as we found that the weighted cross-entropy loss did not further improve on the validation with more training. We then evaluated the model once on the test set. We independently repeated this process five times, using random seeds for the bootstrap sampling, training/validation splits and network weight initialisations. This allowed us to generate a distribution of performance of the DCE, such that we could report a mean and SD of metrics and conduct statistical tests to compare its performance against the board-certified ophthalmologists. We ensured that the model was not given any information about the test set, such as how many samples of each class to expect.

Human expert classification

We asked seven board-certified staff ophthalmologists (mean practice duration: 2.4 years, range: 1–7 years) to independently classify each image in the test set into one of the four predetermined classes (normal, DR, glaucoma, AMD), using only information from the image. We also asked each ophthalmologist whether they were 'confident' or 'not confident' in their classification of each image. The ophthalmologists were not informed about the underlying split of the classes (ie, how many images per class were included in the test set) and were only able to select one of the four classes per image. The task was administered remotely over Google Forms.

Evaluation metrics

Several metrics were measured to compare the performances of the DCE and ophthalmologists. We calculated the overall accuracy defined as the percentage of correct predictions over all test images, as well as the overall (macroaveraged over all four classes) F1-score, positive predictive value (PPV), sensitivity and specificity. We also measured these metrics per class in a one-versus-all manner. We use the conventional definition of F1-score as the harmonic mean of PPV and sensitivity with equal weighting:

$$F_1 = 2\frac{PPV \times Sensitivity}{PPV + Sensitivity}$$

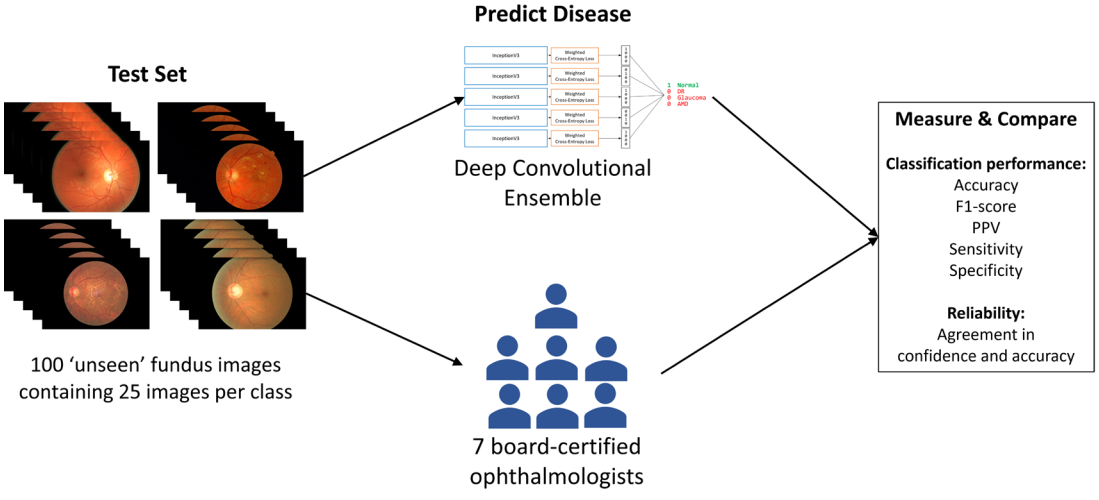

Figure 2 Experiment overview. A test set consisting of 100 images was rated independently by the DCE five times and by each of seven board-certified ophthalmologists. Classification metrics and reliability measures were compared between the DCE and ophthalmologist predictions. AMD, age-related macular degeneration; DCE, deep convolutional ensemble; DR,diabetic retinopathy; PPV, positive predictive value.

We acknowledge that PPV is dependent on the true prevalence of each respective class, which will be different to the 25% in our test set. However, we report the PPV is solely a means of comparing relative performance between the DCE and ophthalmologists. For the DCE, we also report the AUROC averaged over all four classes. It was not possible to report the AUROC for the ophthalmologists as we did not ask the ophthalmologists to report their prediction decisions at multiple confidence levels.

To understand the reliability of predictions, we looked at the agreement between the confidence and accuracy in each prediction by the DCE and ophthalmologists. We would expect a truly reliable classifier to only be confident when it is accurate and not confident when it is inaccurate.[28]

Statistical analyses

We conducted two-sample t-tests, assuming unknown and unequal variances, to determine statistically significant differences in metrics between the DCE and ophthalmologists.

RESULTS

We report the results of classification performance and reliability on the test set experiment. Unless stated otherwise, the order of numerical results below always leads with the DCE followed by the ophthalmologists.

Classification performance

Over all 100 test images and four classes, we found that the DCE had a mean higher overall accuracy than the ophthalmologists (79.2% vs 72.7%, p=0.03), as well as a higher mean overall F1-score (79.9% vs 72.2%, p=0.02), higher mean overall PPV (85.0% vs 77.4%, p=0.0005), higher mean overall sensitivity (79.2% vs 72.7%, p=0.03) and a higher mean overall specificity (93.1% vs 90.9%, p=0.03). Figure 3 illustrates these results as boxplots. The DCE classification performance corresponded to a mean class-averaged AUROC of 0.9424 (SD: 0.0014). A mean of 1.8% (range: 0.0%–3.0%) of response output by the DCE did not constitute a majority vote.

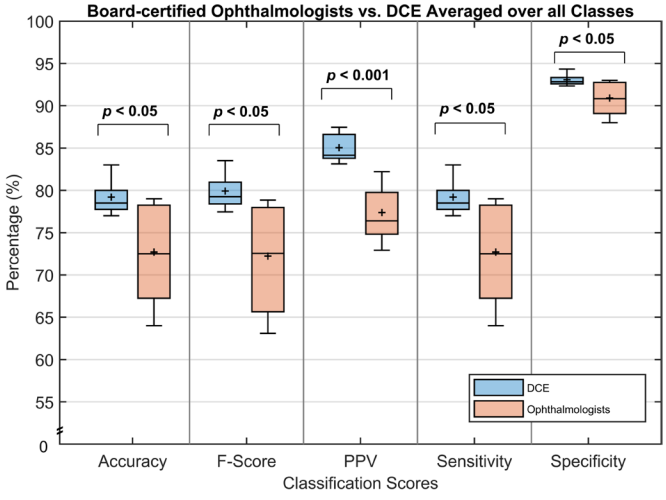

Figure 3 Classification scores for both the DCE and ophthalmologists over all 100 test set images and four classes. Box plots include a horizontal solid line and solid cross indicating the median and mean values, respectively, for each score. P values less than 0.05 are indicated, as determined by a two-sample t-test. DCE, deep convolutional ensemble; PPV, positive predictive value.

In classifying normal fundus images, we found there were no statistically significant differences between the DCE and ophthalmologists in the mean F1-score (73.0% vs 70.5%, p=0.39), mean PPV (59.3% vs 61.3%, p=0.72), mean sensitivity (95.2% vs 87.4%, p=0.07) and the mean specificity (78.1% vs 78.7%, p=0.92).

In classifying DR, the DCE had a statistically significant higher mean F1-score than the ophthalmologists (76.8% vs 57.5%, p=0.01), a statistically higher mean sensitivity (72.8% vs 49.7%, p=0.01), while achieving a similar mean PPV (81.8% vs 73.7%, p=0.18) and mean specificity (94.4% vs 93.7%, p=0.75).

For glaucoma classification, we found no statistically significant differences between the DCE and ophthalmologists. The DCE had a comparable mean F1-score (83.9% vs 75.7%, p=0.10), mean PPV (100% vs 88.9%, p=0.06), mean sensitivity (72.8% vs 68.6%, p=0.58) and mean specificity (100% vs 96.2%, p=0.10).

Lastly, in AMD classification, we found that the DCE had a comparable mean F1-score as the ophthalmologists (85.9% vs 85.2%, p=0.69), a statistically higher mean PPV (99.0% vs 85.6%, p=0.0006), a statistically lower mean sensitivity (76.0% vs 85.1%, p=0.01) and a statistically higher mean specificity (99.7% vs 95.0%, p=0.002). Figure 4 plots the classification performance per class, comparing the DCE and ophthalmologists.

Table 2 provides the confusion matrix for the DCE and the board-certified ophthalmologists, summarising the mean per cent agreement between the predicted class against the ground-truth labels.

Reliability

We found that the DCE had an overall higher mean agreement in confidence and accuracy, compared with the ophthalmologists (81.6% vs 70.3%, p<0.001). Specifically, the DCE was confident when accurate with a higher mean frequency compared with ophthalmologists (77.4% vs 58.7%, $p<10^{-5}$). The DCE was not confident while inaccurate with a lower mean frequency (4.2% vs 11.6%, p=0.001). Conversely, the ophthalmologists had a higher mean frequency of being not confident when accurate (ophthalmologists: 14%, DCE: 1.8%, p=0.002), and both methods had a similar mean frequency of being confident when inaccurate (16.6% vs 15.7%, p=0.80). Table 3 summarises these results. We observed that the DCE had a skewed, unimodal distribution of confidence values, with a mean of 94.0% confidences greater than 0.5 (table 3), and 50% of confidence values greater than 0.807. On the other hand, the board-certified ophthalmologists denoted a mean of 25.6% test images as 'not confident'.

Table 4 provides the confusion matrix of only the 'confident' predictions for both the DCE and board-certified ophthalmologists. This table illustrates the mean per cent agreement between the 'confident' predictions and the ground-truth labels. Figure 5A–C provide examples of fundus photographs that both DCE and ophthalmologists were completely confident in, and one each where the DCE and ophthalmologists were least confident in, as well as their respective diagnoses.

DISCUSSION AND CONCLUSION

We developed an ensemble of deep CNNs which we showed to be more accurate than seven board-certified ophthalmologists at classifying 100 fundus images, both in terms of overall mean accuracy and F1-score over the four image classes. The majority of this difference stems from the DCE's superiority in classifying DR images compared with the ophthalmologists (figure 4), which is statistically significant. We believe this better performance is the result of the DCE's ability to detect mild presentations of DR in fundus images compared with ophthalmologists, as the

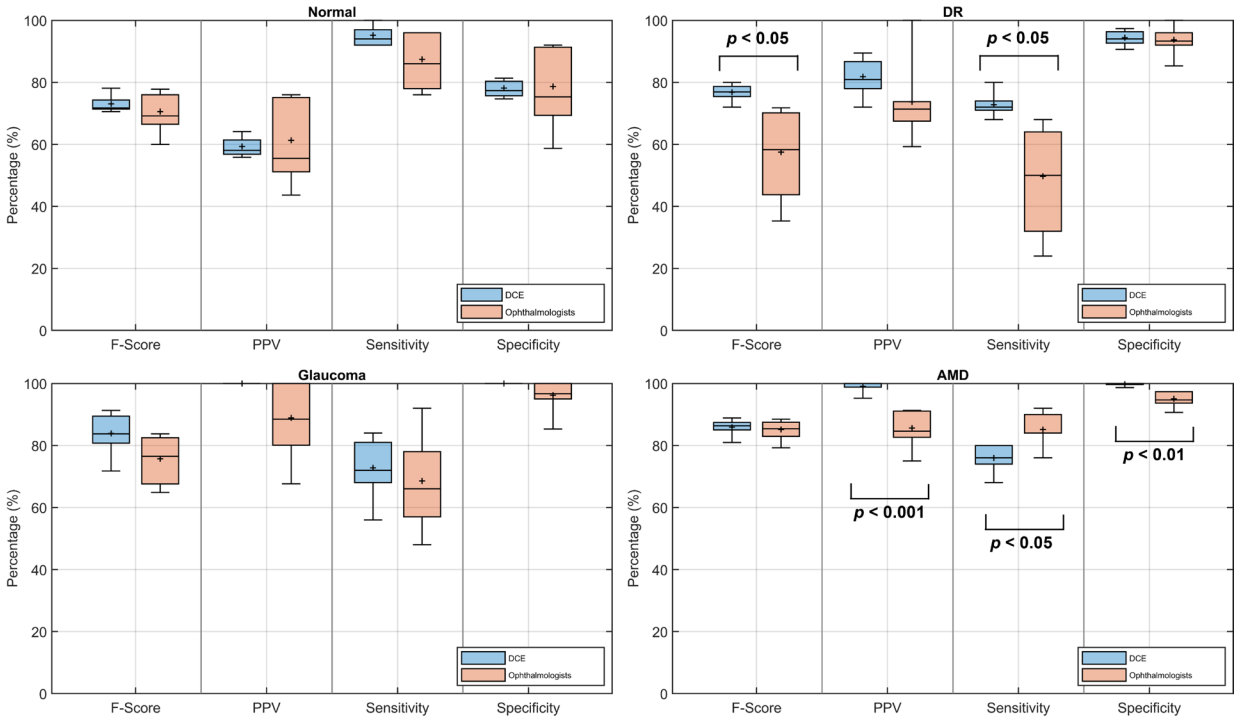

Figure 4 Classification scores for both the DCE and ophthalmologists per class in the test set. Box plots include a horizontal solid line and solid cross indicating the median and mean, values. respectively for each score. P values less than 0.05 are indicated, as determined by a two-sample t-test. AMD, age-related macular degeneration; DCE, deep convolutional ensemble; DR,diabetic retinopathy; PPV, positive predictive value.

datasets the DCE was trained on contained a wide spectrum of DR presentations. On the other hand, ophthalmologists do not detect DR from images alone and would also use a dilated clinical fundus examination to make this diagnosis. We verified this by manually reviewing the images which were incorrectly classified by the majority of ophthalmologists but correctly classified by the DCE and found that the majority of these (54.5%) fundi were classified by the ophthalmologists as 'normal' when they had mild DR. In contrast, the DCE did not exceed the ophthalmologists' performance in classes where the number of training samples and original datasets were limited, such as for glaucoma and AMD. Nevertheless, we found that the DCE exhibited statistically equivalent or superior performance to ophthalmologists in all metrics over all classes, with the exception of sensitivity in AMD detection in which the ophthalmologists achieved a mean score of 85.1% compared with DCE's mean of 76.0% (p=0.01; figure 4). Altogether, these results demonstrate that the DCE model has a higher accuracy in detecting and classifying disease from fundus images alone compared with ophthalmologists. To the best of our knowledge, this is the first study of its kind to show both superior classification performance and reliability compared with ophthalmologists for classifying multiple retinal diseases based on fundus photographs, although similar results have been demonstrated in lung lesion detection in radiographs[29] and in skin lesion detection in photographs.[30]

Our study also found that the DCE was more reliable in its predictions compared with ophthalmologists, as the DCE had a higher mean agreement between its stated confidence and accuracy compared with ophthalmologists. Our analysis showed that this was primarily due to the large proportion of underconfident responses (not confident yet accurate) given by the ophthalmologists compared with the DCE (table 3). As above, this could be explained by the fact ophthalmologists do not recognise pathology purely from fundus photographs but also rely on the dilated retinal examination and

Table 2 Confusion matrices for deep convolutional ensemble and board-certified ophthalmologists showing the mean (and SD) per cent agreement between the predicted labels against the ground-truth labels over the test set

		Deep convolutional ensemble				Ophthalmologists			
		Normal	DR	Glaucoma	AMD	Normal	DR	Glaucoma	AMD
Ground-truth labels	Normal	23.8% (0.8%)	1.2% (0.8%)	0.0% (0.0%)	0.0% (0.0%)	21.9% (2.3%)	1.3% (1.4%)	1.4% (2.1%)	0.4% (0.5%)
	DR	6.6% (1.1%)	18.2% (1.1%)	0% (0.0%)	0.2% (0.4%)	9.7% (5.1%)	12.4% (4.5%)	1.0% (1.4%)	1.9% (1.1%)
	Glaucoma	6.6% (2.3%)	0.2% (0.4%)	18.2% (2.7%)	0.0% (0.0%)	5.6% (4.8%)	0.9% (1.6%)	17.1% (3.8%)	1.4% (0.5%)
	AMD	3.2% (0.4%)	2.8% (1.5%)	0.0% (0.0%)	19% (1.2%)	0.7% (0.8%)	2.6% (1.5%)	0.4% (0.5%)	21.3% (1.4%)

Green cells indicate agreement between the ground-truth labels and predictions by the deep convolutional ensemble or ophthalmologists, and red cells similarly indicate disagreement
AMD, age-related macular degeneration; DR, diabetic retinopathy.

Table 3 Mean (and SD) per cent agreement between confidence and accuracy of the deep convolutional ensemble and ophthalmologists

	Deep convolutional ensemble		Ophthalmologists	
	Correct	Incorrect	Correct	Incorrect
Confident	77.4% (2.5%)	16.6% (2.5%)	58.7% (4.3%)	15.7% (8.6%)
Not confident	1.8% (0.8%)	4.2% (0.8%)	14% (6.0%)	11.6% (3.7%)

Green cells indicate agreement between the ground-truth labels and predictions by the deep convolutional ensemble or ophthalmologists, and red cells similarly indicate disagreement

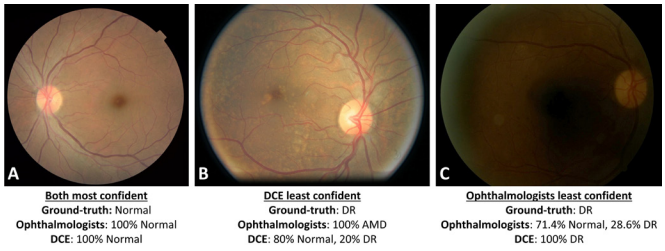

Figure 5 Examples of fundus photographs showing least and most confident predictions. (A:) 'Normal' fundus image predicted with greatest confidence by all five DCE models and all seven ophthalmologists. (B:) 'DR' fundus image with the lowest mean confidence as rated by the DCE. (C:) 'DR' fundus image with the lowest mean confidence as rated by the ophthalmologists. AMD, age-related macular degeneration; DCE, deep convolutional ensemble; DR,diabetic retinopathy.

auxiliary testing (such as OCT and visual fields). Additionally, the test set included fundus photographs of variable quality, many of which would be considered suboptimal for the detection of retinal disease—as evident in figure 5C which demonstrates the image rated least confidently by the ophthalmologists. Both the DCE and ophthalmologists had a similar rate of being overconfident (confident yet inaccurate), confirming that ensembling leads to well-calibrated classification in a manner that is equivalent to or better than human experts.[27 31] A high agreement between confidence and accuracy is promising when considering an algorithm for clinical application, as the confidence values output by a model can be more meaningfully interpreted on newly acquired patient images when the ground-truth pathology is still unknown.

Our test set was limited to 100 fundus images, which is a relatively small sample size for evaluating modern machine learning methods. However, this sample size was chosen so that the ophthalmologists could perform image classification in one session without fatiguing. Another limitation of using previously published image sets for training and testing is the lack of access to clinical data in addition to the fundus photographs. As such, we have assumed that the ground-truth labels are accurate and that fundus photographs contain single diseases only. However, datasets used different criteria to grade retinopathies—for instance DiaretDB relied on ophthalmologists to manually detect visual findings in the fundus photographs to determine the presence of DR,[14] whereas clinical diagnoses were used as ground-truth labels in the MESSIDOR dataset.[21] It was not possible to standardise labels across the data sources, as each institution used different criteria for grading and clinical diagnoses for each eye were not available. It was not possible to guarantee images contained only single diseases for the same reason. This introduces a certain amount of noise, uncertainty and inconsistency in the training and test sets, which the DCE model learns but the board-certified ophthalmologists may not be aware of. Moreover, as our test set was proportionally sampled from the same data sources used in our training/validation pipeline, datasets were under-represented or over-represented in the test set based on the total number of images they contained for each disease category. Because the DCE was trained on the same distribution of data sources, and as some datasets contained a much greater number of certain conditions compared with others, this potentially biased the comparison with board-certified ophthalmologists who were not familiar with the data sets prior to grading the test set. Future work can address these limitations by collecting a prospective multidisease photographic database with associated clinical data.

We further explored the ophthalmologists' responses on the test set to determine whether there were any images for which all ophthalmologists were in disagreement with the prescribed ground-truth label, but also had 100% consensus on the classification. There were two such images, both of which were labelled as DR in the original dataset but were rated as 'normal' and 'AMD', respectively, by all ophthalmologists. Given this consensus, we ran our statistical analysis after removing these two images. We found that the DCE maintained a higher mean accuracy than the ophthalmologists (80.4% vs 74.2%, p=0.04), as well as a higher mean F1-score (81.0% vs 73.7%, p=0.03), over all 98 test set images. The DCE also had statistically higher mean PPV, sensitivity and specificity than the ophthalmologists over all images.

In this study we showed that it is possible to train an ensemble of deep CNNs to accurately identify three retinal pathologies and normal retinas from colour fundus photographs alone. We showed

Table 4 Confusion matrices for deep convolutional ensemble and board-certified ophthalmologists showing the mean (and SD) per cent agreement between 'confident' predicted labels against the ground-truth labels over the test set

		Deep convolutional ensemble				Ophthalmologists			
		Normal	DR	Glaucoma	AMD	Normal	DR	Glaucoma	AMD
Ground-truth labels	Normal	25.1% (0.9%)	1.3% (0.9%)	0.0% (0.0%)	0.0% (0.0%)	22.1% (3.0%)	0.3% (0.6%)	0.2% (0.6%)	0.2% (0.5%)
	DR	5.7% (1.2%)	18.9% (1.0%)	0.0% (0.0%)	0.0% (0.0%)	7.8% (5.5%)	12.7% (6.8%)	0.2% (0.6%)	2.1% (1.1%)
	Glaucoma	5.7% (2.3%)	0.2% (0.5%)	18.9% (2.7%)	0.0% (0.0%)	4.3% (4.1%)	0.4% (1.0%)	18.5% (4.2%)	2.0% (0.8%)
	AMD	2.8% (0.6%)	1.9% (1.2%)	0.0% (0.0%)	19.4% (1.5%)	0.5% (0.9%)	2.1% (1.6%)	0.2% (0.6%)	26.3% (2.3%)

Green cells indicate agreement between the ground-truth labels and predictions by the deep convolutional ensemble or ophthalmologists, and red cells similarly indicate disagreement
AMD, age-related macular degeneration; DR, diabetic retinopathy.

that this performance meets or exceeds the performance of human experts in the field, and further that the reliability (or confidence calibration) is better than that of the board-certified ophthalmologists. Although we use InceptionV3, a previously developed deep learning model, we showed that it is possible to use existing pretrained architectures in an ensemble configuration to meet, or even surpass, human expert medical image classification accuracy and confidence calibration. We expect future avenues of research to explore how technical advancements in model architecture and training algorithms might further advance classification accuracy and reliability of supervised learning algorithms. While clinicians typically have access to additional information such as clinical history, a clinical examination and auxiliary testing to assist with making these diagnoses, these tests are costly in both human and technical resources. Automated artificial intelligence (AI) classifiers could represent a method by which rapid population-based screening for retinal disease could be performed using fundus photographs alone. Future work should explore the potential deployment of multidisease AI classifiers to assist with community-based retinal screening, particularly in settings where access to ophthalmology diagnostics is limited.

Twitter Stephan Ong Tone @StephanOngTone and Jovi C Y Wong @jovicywong

Acknowledgements We would like to thank Dr Aaron Y. Lee (University of Washington) for providing feedback on our study.

Contributors JCYW initiated and planned the project, obtained grant funding, designed the data collection tools, monitored the data collection, implemented the methodology, analysed the data, and drafted and revised the manuscript. JCYW is the guarantor of this study. PUP performed statistical analysis on the data, and drafted and revised the paper. JAM provided feedback on the project plan and methodology, assisted with data collection and revised the manuscript. BGB, PGC, AJK, DJM, SOT and MJW assisted with data collection and revised the manuscript.

Funding This work is funded by the Fighting Blindness Canada Eye on the Cure Grant (JCYW and JAM). JCYW is a 2020 Joule Innovation Grant recipient. The sponsors or funding organisations had no role in the design or conduct of this research.

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. All datasets are publicly available and their sources were cited in the manuscript.

ORCID iDs
Prashant U Pandey http://orcid.org/0000-0001-9275-590X
David J Mathew http://orcid.org/0000-0001-8488-2060
Michael J Wan http://orcid.org/0000-0001-5110-5142

REFERENCES

1 Ting DSW, Pasquale LR, Peng L, *et al*. Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol* 2019;103:167–75.
2 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;316:2402–10.
3 Saba T, Akbar S, Kolivand H, *et al*. Automatic detection of papilledema through fundus retinal images using deep learning. *Microsc Res Tech* 2021;84:3066–77.
4 Phasuk S, Tantibundhit C, Poopresert P, *et al*. Automated glaucoma screening from retinal fundus image using deep learning. *Annu Int Conf IEEE Eng Med Biol Soc* 2019;2019:904–7.
5 Ipp E, Liljenquist D, Bode B, *et al*. Pivotal evaluation of an artificial intelligence system for autonomous detection of referrable and vision-threatening diabetic retinopathy. *JAMA Netw Open* 2021;4:e2134254.
6 Abràmoff MD, Lavin PT, Birch M, *et al*. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med* 2018;1:39.
7 Chen Q, Peng Y, Keenan T, *et al*. A multi-task deep learning model for the classification of age-related macular degeneration. *AMIA Jt Summits Transl Sci Proc* 2019;2019:505–14.
8 Cen L-P, Ji J, Lin J-W, *et al*. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nat Commun* 2021;12:4828.
9 Son J, Shin JY, Kim HD, *et al*. Development and validation of deep learning models for screening multiple abnormal findings in retinal fundus images. *Ophthalmology* 2020;127:85–94.
10 Li F, Wang Y, Xu T, *et al*. Deep learning-based automated detection for diabetic retinopathy and diabetic macular oedema in retinal fundus photographs. *Eye (Lond)* 2022;36:1433–41.
11 Ting DSW, Cheung C-L, Lim G, *et al*. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* 2017;318:2211–23.
12 Han Y, Li W, Liu M, *et al*. Application of an anomaly detection model to screen for ocular diseases using color retinal fundus images: design and evaluation study. *J Med Internet Res* 2021;23:e27822.
13 Zapata MA, Royo-Fibla D, Font O, *et al*. Artificial intelligence to identify retinal fundus images, quality validation, laterality evaluation, macular degeneration, and suspected glaucoma. *Clin Ophthalmol* 2020;14:419–29.
14 Kauppi T, Kalesnykiene V, Kamarainen J-K, *et al*. The DIARETDB1 diabetic retinopathy database and evaluation protocol. British Machine Vision Conference 2007; :15. Warwick.
15 Sivaswamy J, Krishnadas SR, Datt Joshi G, *et al*. Drishti-gs: retinal image dataset for optic nerve head(ONH) segmentation. 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI 2014); Beijing, China.2014:53–6.
16 Staal J, Abràmoff MD, Niemeijer M, *et al*. Ridge-based vessel segmentation in color images of the retina. *IEEE Trans Med Imaging* 2004;23:501–9.
17 Budai A, Bock R, Maier A, *et al*. Robust vessel segmentation in fundus images. *Int J Biomed Imaging* 2013;2013:154860.
18 Porwal P, Pachade S, Kamble R, *et al*. Indian diabetic retinopathy image dataset (idrid). *IEEE Dataport* 2018.
19 Cuadros J, Bresnick G. EyePACS: an adaptable telemedicine system for diabetic retinopathy screening. *J Diabetes Sci Technol* 2009;3:509–16.
20 Li N, Li T, Hu C, *et al*. A benchmark of ocular disease intelligent recognition: one shot for multi-disease detection [Lecture Notes in Computer Science]. In: Wolf F, Gao W, eds. *Benchmarking, Measuring, and Optimizing*. Cham: Springer International Publishing, 2021: 177–93. Available: https://doi.org/ 10.1007/978-3-030-71058-3_11
21 Decencière E, Zhang X, Cazuguel G, *et al*. Feedback on a publicly distributed image database: the messidor database. *Image Anal Stereol* 2014;33:231.
22 Zhang Z, Yin FS, Liu J, *et al*. ORIGA(-light): an online retinal fundus image database for glaucoma analysis and research. *Annu Int Conf IEEE Eng Med Biol Soc* 2010;2010:3065–8.
23 Orlando JI, Fu H, Barbosa Breda J, *et al*. Refuge challenge: a unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Med Image Anal* 2020;59:101570.
24 Hoover A, Kouznetsova V, Goldbaum M. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Trans Med Imaging* 2000;19:203–10.
25 Szegedy C, Vanhoucke V, Ioffe S, *et al*. Rethinking the inception architecture for computer vision. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Vol 2016-Decem. IEEE Computer Society; Las Vegas, NV, USA.2016:2818–26.
26 Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems* 2017;30.
27 Mehrtash A, Wells WM, Tempany CM, *et al*. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Trans Med Imaging* 2020;39:3868–78.
28 Guo C, Pleiss G, Sun Y, *et al*. On calibration of modern neural networks. In: *34th International Conference on Machine Learning, ICML* . 2017: 2017. 2130–43.
29 Rajpurkar P, Irvin J, Ball RL, *et al*. Deep learning for chest radiograph diagnosis: a retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS Med* 2018;15:e1002686.
30 Fujisawa Y, Otomo Y, Ogata Y, *et al*. Deep learning surpasses dermatologists. *Br J Dermatol* 2019;180:e39.
31 Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach Learn* 2021;110:457–506.