Clinically applicable deep learning for diagnosis and referral in retinal disease

Jeffrey De Fauw[1], Joseph R. Ledsam[1], Bernardino Romera-Paredes[1], Stanislav Nikolov[1], Nenad Tomasev[1], Sam Blackwell[1], Harry Askham[1], Xavier Glorot[1], Brendan O'Donoghue[1], Daniel Visentin[1], George van den Driessche[1], Balaji Lakshminarayanan[1], Clemens Meyer[1], Faith Mackinder[1], Simon Bouton[1], Kareem Ayoub[1], Reena Chopra[2], Dominic King[1], Alan Karthikesalingam[1], Cían O. Hughes[1,3], Rosalind Raine[3], Julian Hughes[2], Dawn A. Sim[2], Catherine Egan[2], Adnan Tufail[2], Hugh Montgomery[3], Demis Hassabis[1], Geraint Rees[3], Trevor Back[1], Peng T. Khaw[2], Mustafa Suleyman[1], Julien Cornebise[1,3,4], Pearse A. Keane[2,4]* and Olaf Ronneberger[1,4]*

The volume and complexity of diagnostic imaging is increasing at a pace faster than the availability of human expertise to interpret it. Artificial intelligence has shown great promise in classifying two-dimensional photographs of some common diseases and typically relies on databases of millions of annotated images. Until now, the challenge of reaching the performance of expert clinicians in a real-world clinical pathway with three-dimensional diagnostic scans has remained unsolved. Here, we apply a novel deep learning architecture to a clinically heterogeneous set of three-dimensional optical coherence tomography scans from patients referred to a major eye hospital. We demonstrate performance in making a referral recommendation that reaches or exceeds that of experts on a range of sight-threatening retinal diseases after training on only 14,884 scans. Moreover, we demonstrate that the tissue segmentations produced by our architecture act as a device-independent representation; referral accuracy is maintained when using tissue segmentations from a different type of device. Our work removes previous barriers to wider clinical use without prohibitive training data requirements across multiple pathologies in a real-world setting.

Medical imaging is expanding globally at an unprecedented rate[1,2], leading to an ever-expanding quantity of data that requires human expertise and judgement to interpret and triage. In many clinical specialities there is a relative shortage of this expertise to provide timely diagnosis and referral. For example, in ophthalmology, the widespread availability of optical coherence tomography (OCT) has not been matched by the availability of expert humans to interpret scans and refer patients to the appropriate clinical care[3]. This problem is exacerbated by the marked increase in prevalence of sight-threatening diseases for which OCT is the gold standard of initial assessment[4–7].

Artificial intelligence (AI) provides a promising solution for such medical image interpretation and triage, but despite recent breakthrough studies in which expert-level performance on two-dimensional photographs in preclinical settings has been demonstrated[8,9], prospective clinical application of this technology remains stymied by three key challenges. First, AI (typically trained on hundreds of thousands of examples from one canonical dataset) must generalize to new populations and devices without a substantial loss of performance, and without prohibitive data requirements for retraining. Second, AI tools must be applicable to real-world scans, problems and pathways, and designed for clinical evaluation and deployment. Finally, AI tools must match or exceed the performance of human experts in such real-world situations. Recent work applying AI to

OCT has shown promise in resolving some of these criteria in isolation, but has not yet shown clinical applicability by resolving all three.

Results

Clinical application and AI architecture. We developed our architecture in the challenging context of OCT imaging for ophthalmology. We tested this approach for patient triage in a typical ophthalmology clinical referral pathway, comprising more than 50 common diagnoses for which OCT provides the definitive imaging modality (Supplementary Table 1). OCT is a three-dimensional volumetric medical imaging technique analogous to three-dimensional ultrasonography but measuring the reflection of near-infrared light rather than sound waves at a resolution for living human tissue of $\sim 5\,\mu m$[10]. OCT is now one of the most common imaging procedures with 5.35 million OCT scans performed in the US Medicare population in 2014 alone (see https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier.html). It has been widely adopted across the UK National Health Service (NHS) for comprehensive initial assessment and triage of patients requiring rapid non-elective assessment of acute and chronic sight loss. Rapid access 'virtual' OCT clinics have become the standard of care[11,12]. In such clinics, expert clinicians interpret the OCT and clinical history to diagnose and triage patients with

[1]DeepMind, London, UK. [2]NIHR Biomedical Research Centre at Moorfields Eye Hospital and UCL Institute of Ophthalmology, London, UK. [3]Present address: University College London, London, UK. [4]These authors contributed equally: Julien Cornebise, Pearse A. Keane, Olaf Ronneberger. *e-mail: pearse.keane@moorfields.nhs.uk; olafr@deepmind.com
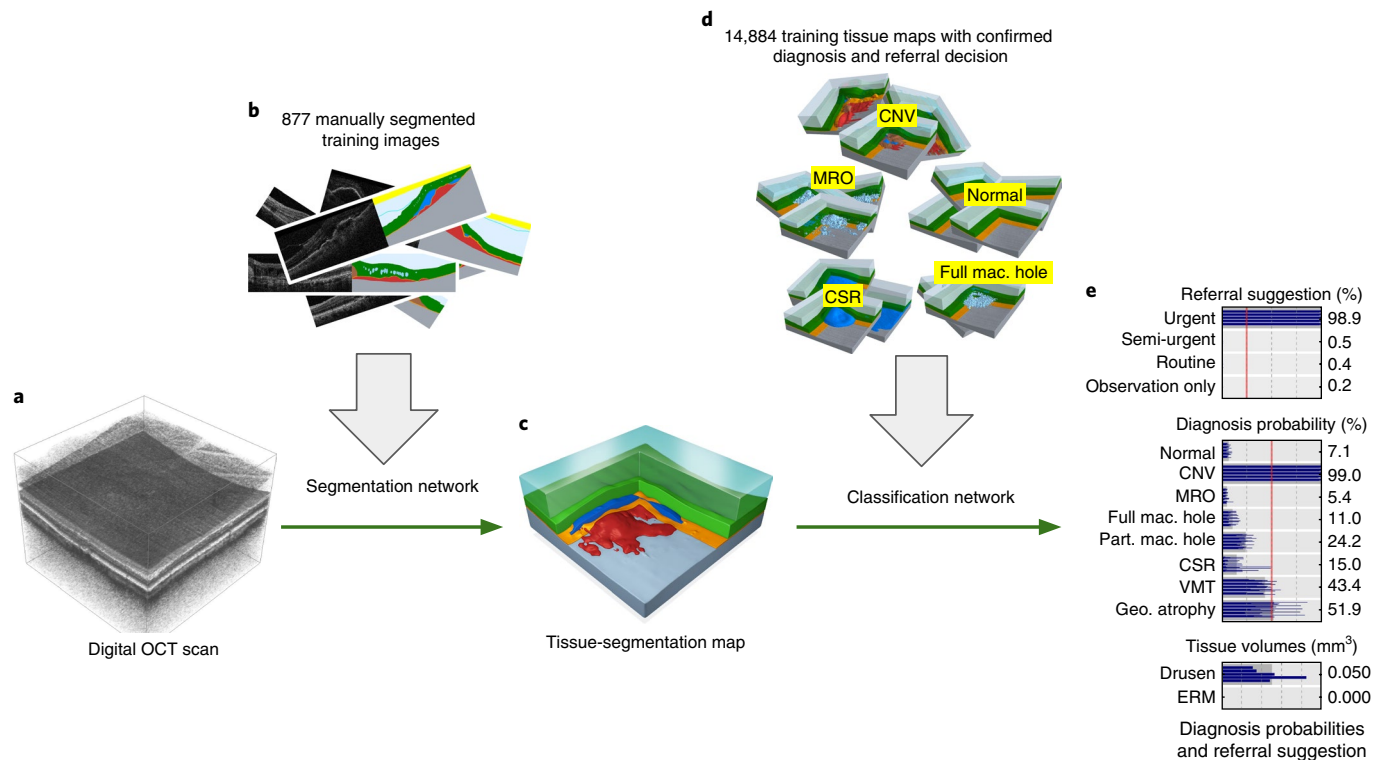
Fig. 1 | Our proposed AI framework. a, Raw retinal OCT scan (6×6×2.3 mm³ around the macula). **b**, Deep segmentation network, trained with manually segmented OCT scans. **c**, Resulting tissue segmentation map. **d**, Deep classification network, trained with tissue maps with confirmed diagnoses and optimal referral decisions. **e**, Predicted diagnosis probabilities and referral suggestions.

pathology affecting the macula, the central part of the retina that is required for high-resolution, color vision.

Automated diagnosis of a medical image, even for a single disease, faces two main challenges: technical variations in the imaging process (different devices, noise, ageing of the components and so on), and patient-to-patient variability in pathological manifestations of disease. Existing deep learning approaches[8,9] tried to deal with all combinations of these variations using a single end-to-end black-box network, thus typically requiring millions of labeled scans. By contrast, our framework decouples the two problems (technical variations in the imaging process and pathology variants) and solves them independently (see Fig. 1). A deep segmentation network (Fig. 1b) creates a detailed device-independent tissue-segmentation map. Subsequently, a deep classification network (Fig. 1d) analyses this segmentation map and provides diagnoses and referral suggestions.

The segmentation network (Fig. 1b) uses a three-dimensional U-Net architecture[13,14] to translate the raw OCT scan into a tissue map (Fig. 1c) with 15 classes including anatomy, pathology and image artefacts (Supplementary Table 2). It was trained with 877 clinical OCT scans (Topcon 3D OCT, Topcon) with sparse manual segmentations (dataset 1 in Supplementary Table 3, see Methods 'Manual segmentation' and 'Datasets' for full breakdown of scan dataset). Only approximately three representative slices out of the 128 slices of each scan were manually segmented (see Supplementary Table 4 for image sizes). This sparse annotation procedure[14] allowed us to cover a large variety of scans and pathologies with the same workload as approximately 21 dense manual segmentations. Examples of the output of our segmentation network for illustrative pathologies are shown in Fig. 2.

The classification network (Fig. 1d) analyses the tissue-segmentation map (Fig. 1c) and as the primary outcome provides one of four referral suggestions currently used in clinical practice at Moorfields Eye Hospital (please see Supplementary Table 1 for a

list of retinal conditions associated with these referral suggestions). Additionally, it reports the presence or absence of multiple, concomitant retinal pathologies (Supplementary Table 5). To construct the training set for this network, we assembled 14,884 OCT scan volumes obtained from 7,621 patients who were referred to the hospital with symptoms suggestive of macular pathology (see Methods 'Clinical labeling'). These OCT scans were automatically segmented using our segmentation network. The resulting segmentation maps with the clinical labels built the training set for the classification network (dataset 3 in Supplementary Table 3, illustrated in Fig. 1d).

A central challenge in OCT-image segmentation is the presence of ambiguous regions, where the true tissue type cannot be deduced from the image, and thus multiple equally plausible interpretations exist. To address this issue, we trained not one but multiple instances of the segmentation network. Each network instance creates a full segmentation map for the given scan, resulting in multiple hypotheses (see Supplementary Fig. 1). Analogous to multiple human experts, these segmentation maps agree in areas with clear image structures but may contain different (but plausible) interpretations in ambiguous low-quality regions. These multiple segmentation hypotheses from our network can be displayed as a video, in which the ambiguous regions and the proposed interpretations are clearly visible (see Methods 'Visualization of results in clinical practice'; use of this viewer across a range of challenging macular diseases is illustrated in Supplementary Videos 1–9).

Achieving expert performance on referral decisions. To evaluate our framework, we first defined a gold standard. This used information that is not available at the first patient visit and OCT scan, by examining the patient clinical records to determine the final diagnosis and optimal referral pathway in the light of the (subsequently obtained) information. Such a gold standard can only be obtained retrospectively. Gold standard labels were acquired for
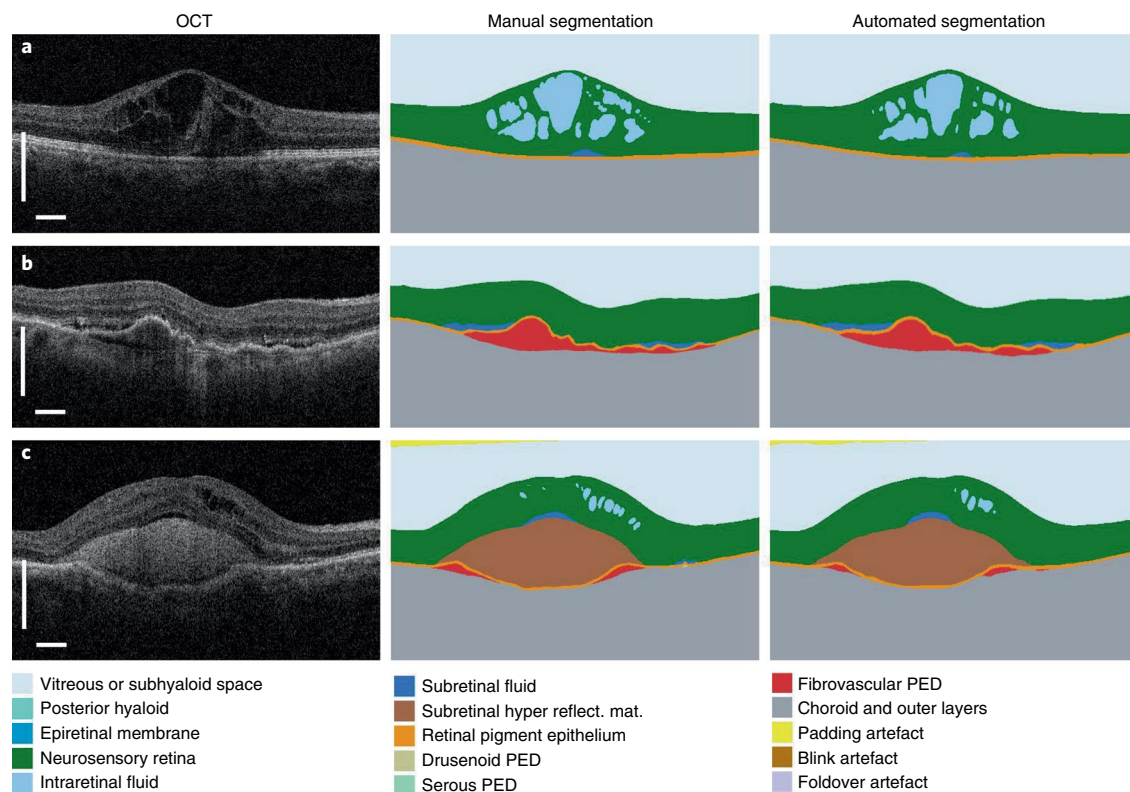
OCT   Manual segmentation   Automated segmentation

Vitreous or subhyaloid space   ■ Subretinal fluid   ■ Fibrovascular PED
Posterior hyaloid   ■ Subretinal hyper reflect. mat. ■ Choroid and outer layers
■ Epiretinal membrane   ■ Retinal pigment epithelium   Padding artefact
■ Neurosensory retina   Drusenoid PED   ■ Blink artefact
Intraretinal fluid   Serous PED   Foldover artefact

Fig. 2 | Results of the segmentation network. Three selected two-dimensional slices from the $n = 224$ OCT scans in the segmentation test set (left) with manual segmentation (middle) and automated segmentation (right; detailed color legend in Supplementary Table 2). **a**, A patient with diabetic macular edema. **b**, A patient with choroidal neovascularization resulting from age-related macular degeneration (AMD), demonstrating extensive fibrovascular pigment epithelium detachment and associated subretinal fluid. **c**, A patient with neovascular AMD with extensive subretinal hyperreflective material. Further examples of the variation of pathology with model segmentation and diagnostic performance can be found in Supplementary Videos 1–9. In all examples the classification network predicted the correct diagnosis. Scale bars, 0.5 mm.

997 patients that were not included in the training dataset (dataset 5 in Supplementary Table 5). We then tested our framework on this dataset. For each patient, we obtained the referral suggestion of our framework plus an independent referral suggestion from eight clinical experts, four of whom were retina specialists and four optometrists trained in medical retina (see Supplementary Table 6 for more information). Each expert provided two separate decisions, one (like our framework) from the OCT scan alone (dataset 7 in Supplementary Table 5); and one from the OCT plus fundus image and clinical notes (dataset 8 in Supplementary Table 5, see Supplementary Fig. 2), in two separate sessions spaced at least two weeks apart. We compared each of these performances (framework and two expert decisions) against the gold standard.

Our framework achieved and in some cases exceeded expert performance (Fig. 3). To illustrate this, Fig. 3a displays performance on 'urgent referrals', the most important clinical referral decision (mainly for pathologies that cause choroidal neovascularization; see Supplementary Table 1) versus all other referral decisions as a receiver operating characteristic (ROC) plot (plots for the other decisions are shown in Supplementary Fig. 3). Performance of our framework matched our two best retina specialists and had a significantly higher performance than the other two retinal specialists and all four optometrists when they used only the OCT scans to make their referral suggestion (Fig. 3a, filled markers). When experts had access to the fundus image and patient summary notes to make their decision, their performance improved (Fig. 3a, empty markers) but our framework remained as good as the five best experts and continued to significantly outperform the other three (see Supplementary Information).

To provide a more complete picture, the overall performance of our framework on all four clinical referral suggestions (urgent, semi-urgent, routine and observation only) compared to the two highest performing retina specialists is displayed in Fig. 3b. The framework performed comparably to the two best-performing retina specialists, and made no clinically-serious wrong decisions (top right element of each matrix; that is, referring a patient who needs an urgent referral to observation only). Confusion matrices for the assessments of the other human experts are shown in Supplementary Fig. 4. The aggregated number of wrong referral decisions is displayed as error rate (1 − accuracy) for our framework and all experts in Fig. 3c. Our framework (5.5% error rate) performed comparably to the two best retina specialists (6.7% and 6.8% error rate) and significantly outperformed the other six experts in the 'OCT only' setting. Significance thresholds (3.9% for higher performance and 7.3% for lower performance) were derived by a two-sided exact binomial test, incorporating uncertainty from both the experts and the algorithm (see Methods 'Statistical analysis'). When experts additionally used the fundus image and the summary notes of the patient, five approached the performance of our framework (three retina specialists and two optometrists), which continued to significantly outperform the remaining three (one retina specialist and two optometrists).

Our framework uses an ensemble of five segmentation and five classification model instances (see Supplementary Fig. 1) to achieve these results. Beside the benefits of an uncertainty measure, ensembling also significantly improves overall performance compared to a single model instance. Error rates for different ensemble sizes are shown in Supplementary Fig. 5. With more segmentation
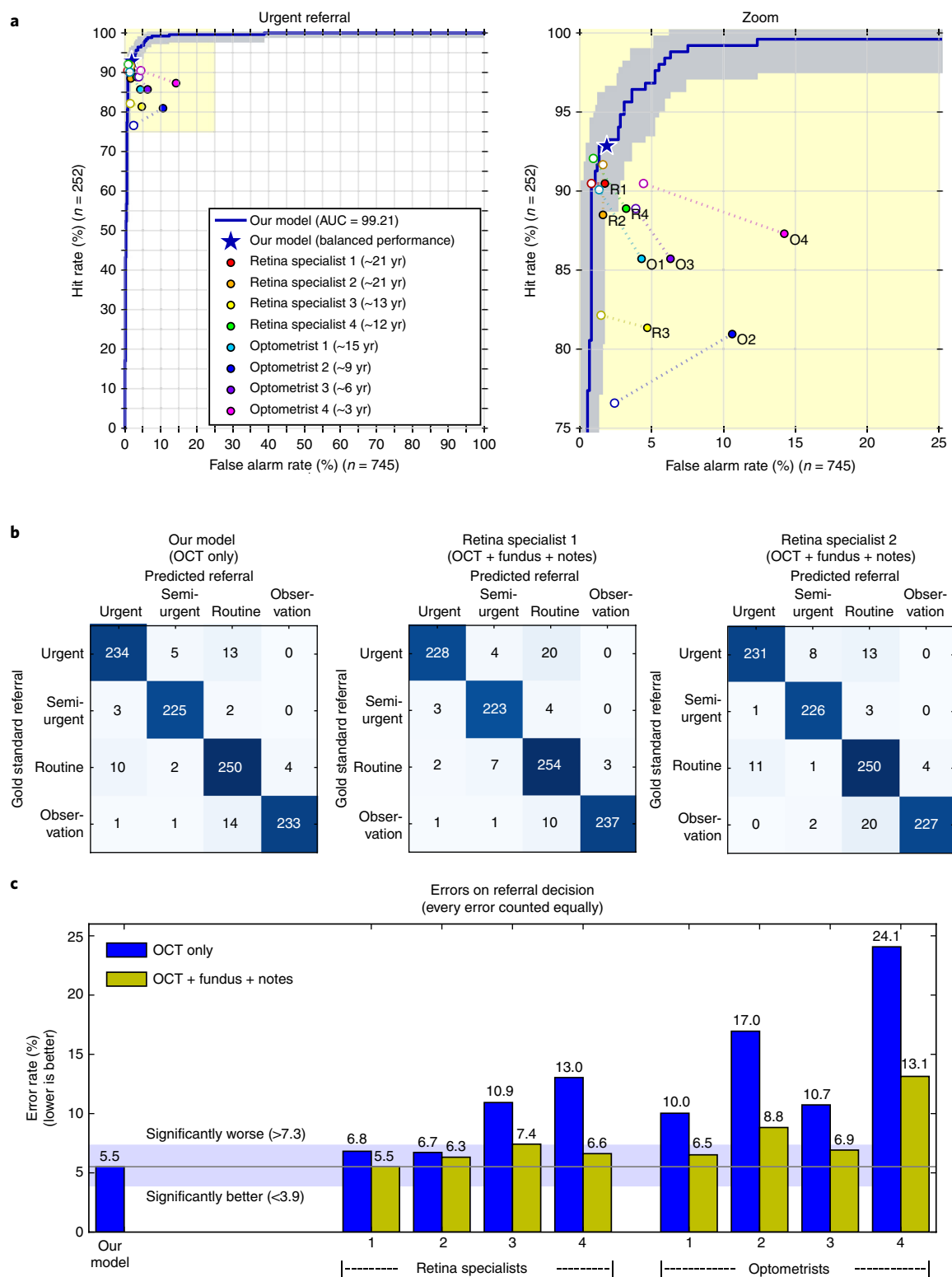
Fig. 3 | Results on the patient referral decision. Performance on an independent test set of $n = 997$ patients (252 urgent, 230 semi-urgent, 266 routine, 249 observation only). **a**, ROC diagram for urgent referral (for choroidal neovascularization (CNV)) versus all other referrals. The blue ROC curve is created by sweeping a threshold over the predicted probability of a particular clinical diagnosis. Points outside the light blue area correspond to a significantly different performance (95% confidence level, using a two-sided exact binomial test). The asterisk denotes the performance of our model in the 'balanced performance' setting. Filled markers denote experts' performance using OCT only; empty markers denote their performance using OCT, fundus image and summary notes. Dashed lines connect the two performance points of each expert. **b**, Confusion matrices with patient numbers for referral decision for our framework and the two best retina specialists. These show the number of patients for each combination of gold standard decision and predicted decision. The numbers of correct decisions are found on the diagonal. Wrong decisions due to overdiagnosis are in the bottom-left triangle, and wrong decisions due to underdiagnosis are in the top-right triangle. **c**, Total error rate (1 – accuracy) on referral decision. Values outside the light-blue area (3.9–7.3%) are significantly different (95% confidence interval, using a two-sided exact binomial test) to the framework performance (5.5%). AUC, area under curve.
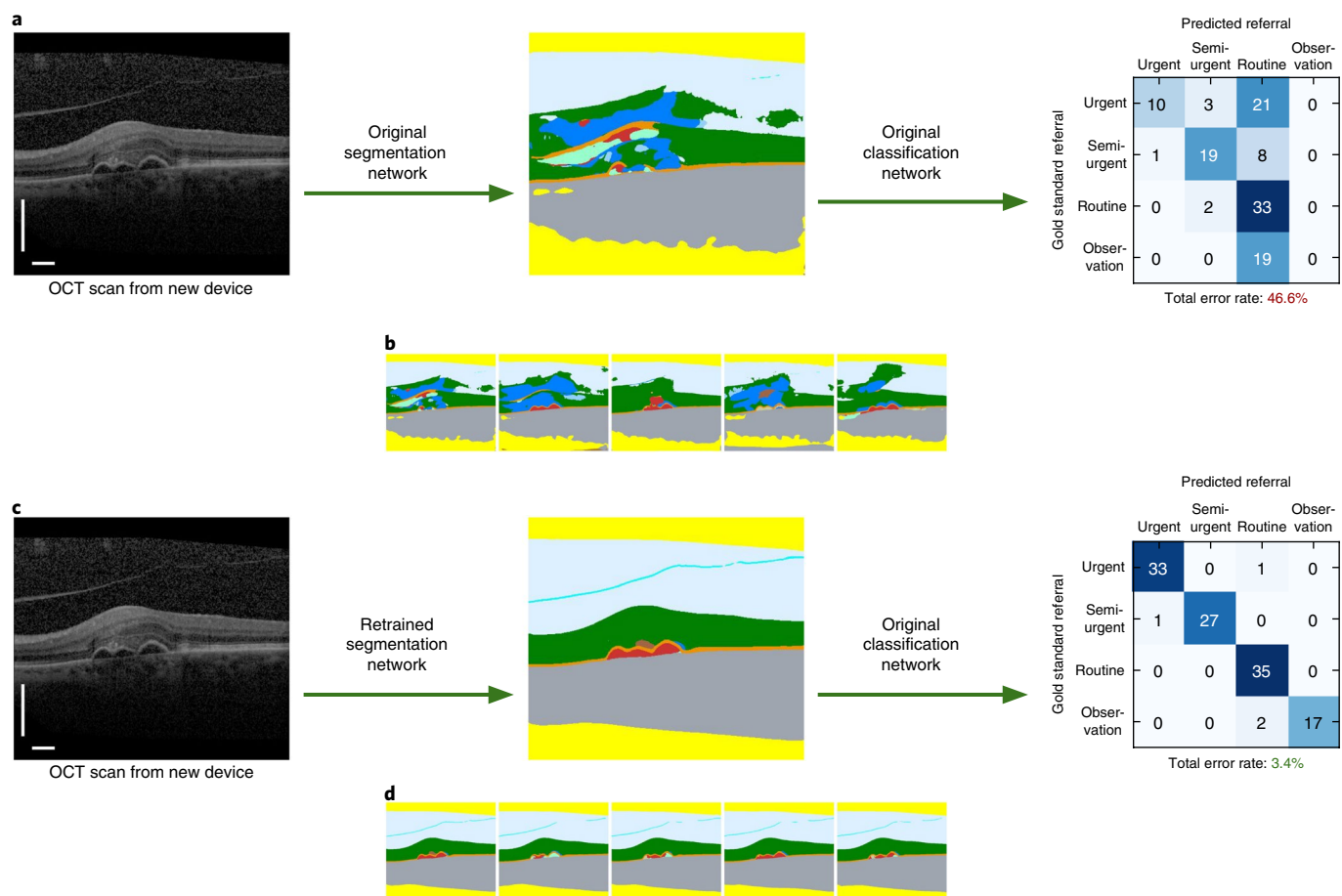
Fig. 4 | Generalization to a new scanning device type. a, Low performance of original network on OCT scans from the new device type 2. Left, the selected slice shows the different appearance of structures in device type 2. Middle, a poor quality segmentation map created with our original segmentation network (color legend in Supplementary Table 2). Right, resulting performance on a new test set of $n = 116$ patients. The confusion matrix shows patient numbers for the referral suggestion. **b**, All five segmentation hypotheses from our original network. The strong variations show the large uncertainty. **c**, High performance was attained on the device type 2 test set ($n = 116$) after retraining the segmentation network with OCT scans from device type 1 and device type 2. The classification network is unchanged. **d**, All five segmentation hypotheses from the retrained segmentation network. The network is confident in the interpretation of most structures, and just highlights the ambiguities in the sub-retinal pigment epithelium (RPE) space. Scale bars: 0.5 mm.

model instances and more classification model instances, performance increases. The bottom right cells in that table illustrate that performance differences between 4×4 model instances and 5×5 model instances are only marginal, so we do not expect significant changes by adding more instances. The accumulated number of diagnostic errors does not fully reflect the clinical consequences that an incorrect referral decision might have for patients, which depends also on the specific diagnosis that was missed. For example, failing to diagnose sight-threatening conditions could result in rapid visual loss[3,15,16], which is not the case for many other diagnoses. For an initial quantitative estimation of these consequences, we weighted different types of diagnostic errors according to the judgement of our clinical experts of the clinical impact of erroneous classification (expressed as penalty points; see Supplementary Fig. 6a). We derived a score for our framework and each expert as a weighted average of all wrong diagnoses. This revealed that our framework achieved a lower average penalty score than any of our experts (Supplementary Fig. 6b). We further optimized the decisions of our framework to minimize this specific score (see Methods 'Optimizing the ensemble output for sensitivity, specificity and penalty scores') which further improved performance (Supplementary Fig. 6b). Therefore, expert performance of our framework is not achieved at the cost of missing clinically important sight-threatening diagnoses.

To examine how our proposed two-stage architecture compared to a traditional single-stage architecture, we trained an end-to-end classification network with the same architecture as our second stage to directly map from a raw OCT scan to a referral decision (see Methods 'End-to-end classification network'). The error rate achieved with an ensemble of five network instances was 5.5%, which was not significantly different from the performance of the two-stage architecture. This validates our choice of the two-stage architecture that offers several clinical advantages (see Supplementary Fig. 7).

Achieving expert performance on retinal morphology. The referral decision recommended by our framework is determined by the most urgent diagnosis detected on each scan (Supplementary Table 1). Patients may also have multiple concomitant retinal pathologies. These additional pathologies do not change the referral decision, but may have implications for further investigations and treatment. Our framework was therefore also trained to predict the probability of a patient having one or more of several pathologies (Supplementary Table 5).

To evaluate performance on diagnosing multiple pathologies, a 'silver standard' for each scan was established by majority vote from the eight experts who evaluated the OCT scan, fundus image and patient summary notes (dataset 6 in Supplementary Table 3). This majority vote biases the assessment against our framework.
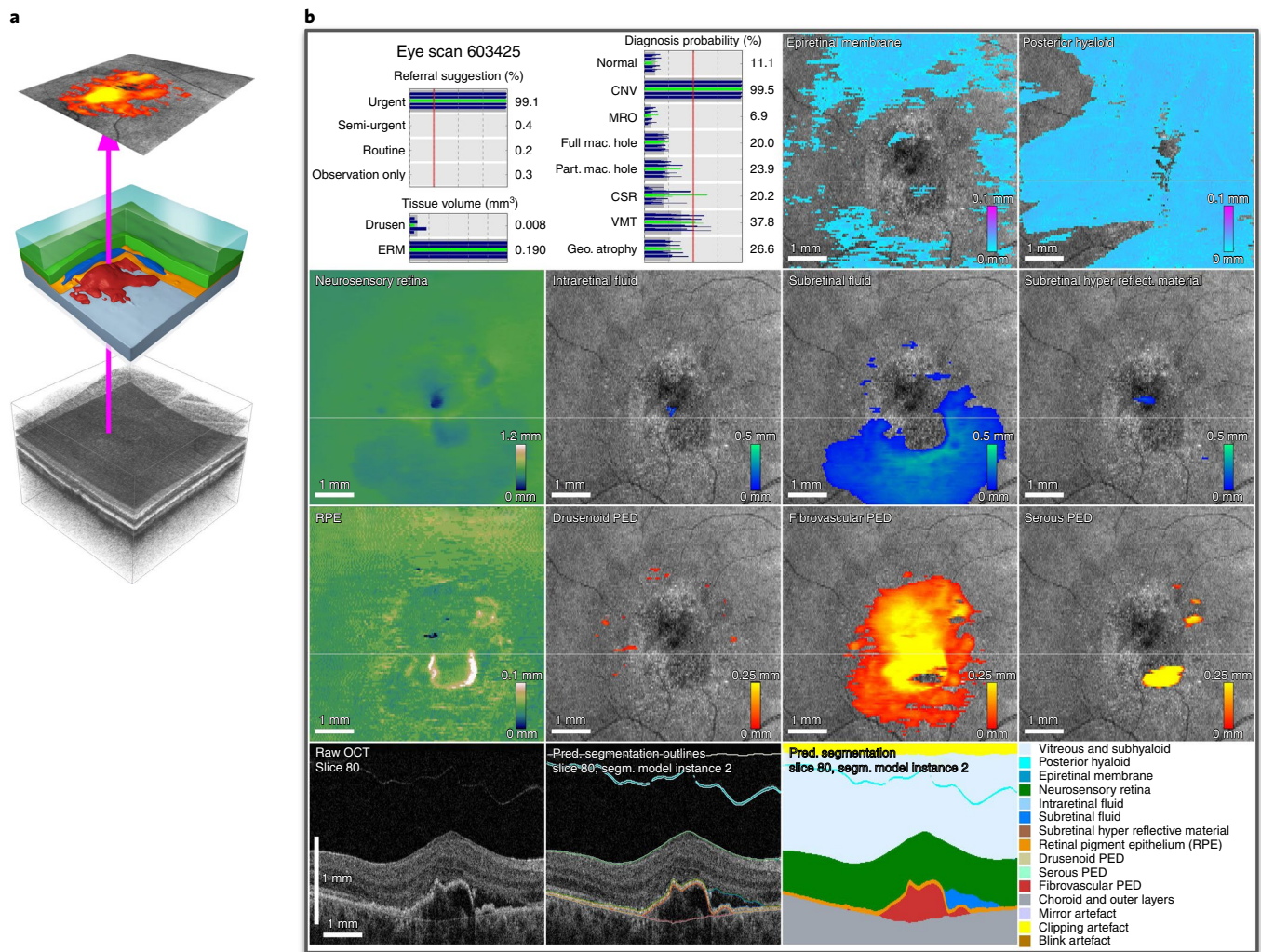
Fig. 5 | Visualization of the segmentation results as thickness maps. a, The average intensity projection of the OCT scan along A-scan direction (frontal view of the eye) is overlaid with a thickness map of the fibrovascular pigment epithelium detachment (PED, red segment). **b**, Screenshot from our OCT viewer. First row (left), referral suggestion, tissue volumes and diagnosis probabilities. The highlighted bars correspond to the selected segmentation model. First–third rows, thickness maps of the 10 relevant tissue types from segmentation model instance 2. The two healthy tissue types (high level retina and RPE) are displayed in a black–blue–green–brown–white color map, the pathological tissues (all others) are displayed as overlay on a projection of the raw OCT scan. The thin white line indicates the position of slice 80. Fourth row, slice 80 from the OCT scan and the segmentation map from segmentation model instance 2. Detailed tissue legend in Supplementary Table 2. The slice and model instance can be interactively selected (see Supplementary Video 1).

Nevertheless, our framework demonstrated an area under the ROC curve that was over 99% for most of the pathologies (and over 96% for all of them; Supplementary Table 7), on par with the performance of the experts on OCT only. As with earlier evaluations, performance of the experts improved when they were provided also with the fundus image and patient summary notes. This improvement was most marked in pathologies classed as 'routine referral', for example geographic atrophy and central serous retinopathy. Many of these pathologies are conditions for which the fundus photograph or demographic information would be expected to provide important information, indicating that there is scope for future work to improve the model. However even in the worst case our framework still performed on par with at least one retinal specialist and one optometrist (Supplementary Table 6 and Supplementary Fig. 8).

Generalization to a new scanning device type. A key benefit of our two-stage framework is the device independence of the second stage. Using our framework on a new device generation thus only requires retraining of the segmentation stage to learn how each

tissue type appears in the new scan, whereas the knowledge about patient-to-patient variability in pathological manifestation of different diseases, which it had learned from the approximately 15,000 training cases, can be reused. To demonstrate this generalization, we collected an independent test set of clinical scans from 116 patients (plus confirmed clinical outcomes) recorded with a different OCT scanner type from a different vendor (Spectralis, Heidelberg Engineering,; hereafter 'device type 2'). This dataset is listed as dataset 11 in Supplementary Table 3 (see Methods 'Datasets'). We selected this device type for several reasons. It is the second most used device type at Moorfields Eye hospital for these examinations, giving rise to a sufficient number of scans. It has a similar worldwide market share as device type 1. But most importantly, this device type provides a large difference in scan characteristics compared to the original device type (see Supplementary Fig. 9).

To evaluate the effect of a different scanning device type, we initially fed the OCT scans from device type 2 into our framework, which was trained only on scans from device type 1 (Fig. 4a). The segmentation network was clearly 'confused' by the changed appearance of these structures and attempted to explain them as

additional retinal layers (Fig. 4a, middle). Consequently, performance was poor with a total error rate for referral suggestions of 46.6% (Fig. 4a, right). Uncertainty of the segmentation network on these (never seen) types of images resulted in five strongly different segmentation hypotheses (Fig. 4b).

We next collected an additional segmentation training set with 152 scans (527 manually segmented slices in total) from this device (dataset 9 in Supplementary Table 3), and retrained the segmentation network with both the training scans from the original device type 1 and the new device type 2 (see Methods 'Segmentation network'). The classification network was not modified.

Our retrained system (adapted segmentation network and unchanged classification network) now achieved a similarly high level of performance on device type 2 as on the original device (Fig. 4c). It suggested incorrect referral decisions for 4 out of the 116 cases, a total error rate of 3.4%. Owing to the small number of cases in the new test set, this is not significantly different from the error rate of 5.5% on device type 1 (P(4 out of 116 < 55 out of 997) = 0.774; see Methods 'Statistical analysis'). For continuity with our previous evaluation, we also measured performance against retina specialists accessing OCT scans plus fundus images and clinical notes (dataset 12 in Supplementary Table 3). Our experts achieved the following error rates (all with access to imaging and clinical notes): retinal specialist one: 2 errors = 1.7% error rate; retinal specialist two: 2 errors = 1.7% error rate; retinal specialist three: 4 errors = 3.4% error rate; retinal specialist four: 3 errors = 2.6% error rate; retinal specialist five: 3 errors = 2.6% error rate. These differences in performance between our framework and the best human retina specialists did not reach statistical significance (P(4 out of 116 > 2 out of 116) = 0.776).

To verify that device type 2 provides the greatest difference in scan characteristics, we performed a feasibility study on the small number of OCT scans from Cirrus HD-OCT 5000 with AngioPlex (Carl Zeiss Meditec) devices available in Moorfields Eye Hospital (dataset of 61 scans; not included here). Applying our original network to these images, we already obtained an error rate of 16.4%. This rate was much lower than that originally obtained with device type 2 (46.6%), consistent with the claim that device type 2 provides a larger difference in scan characteristics from device type 1. Retraining of the segmentation network with 6 manually segmented scans reduced the error rate to 9.8%.

Table 1 summarizes our results. For device type 1, our architecture required 877 training scans with manual segmentations and 14,884 training scans with gold standard referral decisions to achieve expert performance on referral decisions (5.5% error rate). For device type 2, we only required 152 additional training scans with manual segmentations and not a single additional training scan with gold standard referral decisions to achieve the same performance on referral decisions on this device type (3.4% error rate).

Discussion

Recent work in which AI is used for the automated diagnosis of OCT scans shows encouraging results; however, until now such studies have relied on selective and clinically unrepresentative OCT datasets. For example, several authors[17–21] report high performance on automated classification of age-related macular degeneration (AMD) from OCT scans. However, they tested their algorithms on smaller datasets that exclude other pathologies. By contrast, here we demonstrate expert performance on multiple clinical referral suggestions for two independent test datasets of 997 and 116 clinical OCT scans that include a wide range of retinal pathologies.

Several recent studies used deep learning-based architectures to deliver successful segmentation of OCT scans[22–25]. This earlier work focused on a subset of diagnostically relevant tissues types (for example, intraretinal fluid) and applied two-dimensional models in samples of between 10 and 42 patients. In the present work, we go beyond these earlier studies by applying three-dimensional

Table 1 | Number of training scans and achieved performance on the two device types

	Training scans with sparse manual segmentations	Training scans with gold standard referral decision	Test performance on referral decision (error rate)	Test performance on urgent referral (AUC)
Device type 1	877	14,884	55 out of 997 (5.5%)	99.21
Device type 2	152(+877 scans from device type 1)	0	4 out of 116 (3.4%)	99.93

models, segmenting a much larger range of diagnostically relevant tissue types, and connect such segmentation to clinically relevant real-world referral recommendations.

We evaluated our framework on a broad range of real-world images from routine clinical practice at 32 different Moorfields Eye Hospital sites, which cover diverse populations within London and surrounding areas, using 37 individual OCT devices (28 device type 1 and 9 device type 2). The two device types that we tested are both used widely in routine clinical practice at Moorfields Eye Hospital, the largest eye hospital in Europe and North America, and provided a large difference in scan characteristics.

Our framework has a number of potential benefits. The derivation of device-independent segmentation of the OCT scan creates an intermediate representation that is readily viewable by a clinical expert and integrates into clinical workflows (see Fig. 5 for the clinical results viewer). Moreover, the use of an ensemble of five segmentation network instances allows us to present ambiguities arising from the imaging process to the decision network (and could potentially be used for automated quality control).

The 'black box' problem has been identified as an impediment to the application of deep learning in healthcare[26]. Here we created a framework with a structure that closely matches the clinical decision-making process, separating judgements about the scan itself from the subsequent referral decision. This allows a clinician to inspect and visualize an interpretable segmentation, rather than simply being presented with a diagnosis and referral suggestion. Such an approach to medical imaging AI offers potential insights into the decision process, in a fashion more typical of clinical practice. For example, an interpretable representation is particularly useful in difficult and ambiguous cases. Such cases are common in medicine and even expert medical practitioners can find it difficult to reach consensus (for example, our eight experts only agreed on 63.5% of cases even when accessing all information).

Our segmentation map assigns only one label per pixel, and it may not be possible to use the framework directly in other clinical pathways for which the tissue-segmentation map does not contain all required information for a diagnosis (for example, in certain radiomics applications). To keep the advantages of the intermediate device-independent representation in such applications, future work can potentially augment the tissue-segmentation map with multiple labels per pixel to encode local tissue features, or with additional channels that encode continuous features such as an inflammatory reaction. This may be of particular value for other components of the retina, such as the nerve fibre layer, and may be of importance for multiple ocular and brain disorders, such as glaucoma and dementia.

Although we have demonstrated the performance of our framework in the domain of a clinical treatment pathway, the approach has potential utility in clinical training in which the medical professionals must learn to read medical images. In addition, a wide

variety of non-medically qualified health professionals have an interest in appropriately reading and understanding medical images. Our framework produces a visualizable segmentation and achieves expert performance on diagnosis and referral decisions for a large number of scans and pathologies. This therefore raises the intriguing possibility that such a framework could be evaluated as a tool for effectively training healthcare professionals to expert levels.

The segmentation output itself can also be used to quantify retinal morphology and derive measurements of particular pathologies (for example, the location and volume of fibrovascular pigment epithelium detachment and macular edema). Some of these measurements (such as retinal thickness and intraretinal fluid) can currently be derived automatically[27,28], used to investigate correlations with visual outcomes[27] and as an end point in clinical trials of therapies for retinal disease[29–32]. Our framework can be used to define and validate a broader range of automatically derived quantitative measurements.

Our framework can triage scans at first presentation of a patient into a small number of pathways used in routine clinical practice with a performance matching or exceeding both the expert retina specialists and optometrists who staff virtual clinics in a UK NHS setting. Future work can now directly seek evidence for the efficacy of such a framework in a randomized controlled trial. The output of our framework can be optimized to penalize different diagnostic errors, and thus for other clinically important metrics. For example, the potential improvement to patient quality of life of different diagnostic decisions, or avoiding the harm of unnecessary investigation that might come from a false-positive diagnosis, could all be incorporated into future work.

Globally, ophthalmology clinical referral pathways vary, and the range of diseases that can potentially be diagnosed by OCT includes pathologies additional to the macular diseases that were studied here. We studied a major clinical referral pathway in a global center of clinical excellence focusing on 53 key diagnoses relevant to the national (NHS) referral pathways. Our work opens up the possibility of testing the clinical applicability of this approach in other global settings and clinical pathways, such as emergency macular assessment clinics in the UK NHS, triage and assessment in community eye care centers and the monitoring of disease during treatment regimes. Furthermore, devices such as binocular OCT[33] have the potential to increase accessibility in emerging economies. Images produced by such devices will differ in resolution, contrast and image quality from the state-of-the-art devices studied here, and existing AI models trained on current state-of-the-art devices may perform poorly on such new devices. Our proposed two-stage model offers exciting possibilities that enable the use of models more efficiently in countries where state-of-the-art OCT devices are too costly for widespread adoption.

In conclusion, we present a novel framework that analyses clinical OCT scans and makes referral suggestions to a standard that is comparable to clinical experts. Although we focussed on one common type of medical imaging, future work can address a much wider range of medical imaging techniques, and incorporate clinical diagnoses and tissue types well outside the immediate application that was demonstrated here.

References

1. OECD. Computed tomography (CT) exams (indicator). (2017); https://doi.org/10.1787/3c994537-en
2. OECD. Magnetic resonance imaging (MRI) exams (indicator). (2017). https://doi.org/10.1787/1d89353f-en
3. Foot, B. & MacEwen, C. Surveillance of sight loss due to delay in ophthalmic treatment or review: frequency, cause and outcome. *Eye* **31**, 771–775 (2017).
4. Owen, C. G. et al. The estimated prevalence and incidence of late stage age related macular degeneration in the UK. *Br. J. Ophthalmol.* **96**, 752–756 (2012).
5. Rudnicka, A. R. et al. Incidence of late-stage age-related macular degeneration in American whites: systematic review and meta-analysis. *Am. J. Ophthalmol.* **160**, 85–93 (2015).
6. Bourne, R. R. A. et al. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis. *Lancet Glob. Health* **5**, e888–e897 (2017).
7. Schmidt-Erfurth, U., Klimscha, S., Waldstein, S. M. & Bogunović, H. A view of the current and future role of optical coherence tomography in the management of age-related macular degeneration. *Eye* **31**, 26–44 (2017).
8. Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *J. Am. Med. Assoc.* **316**, 2402–2410 (2016).
9. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115––118 (2017).
10. Huang, D. et al. Optical coherence tomography. *Science* **254**, 1178–1181 (1991).
11. Buchan, J. C. et al. How to defuse a demographic time bomb: the way forward? *Eye* **31**, 1519–1522 (2017).
12. Whited, J. D. et al. A modeled economic analysis of a digital teleophthalmology system as used by three federal healthcare agencies for detecting proliferative diabetic retinopathy. *Telemed. J. E Health* **11**, 641–651 (2005).
13. Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. in Navab N., Hornegger J., Wells W., Frangi A. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science, vol. 9351 (Springer, Cham, Switzerland, 2015).
14. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. in Ourselin, S., Joskowicz, L., Sabuncu, M., Unal, G., Wells, W. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. MICCAI 2016. Lecture Notes in Computer Science, vol. 9901 (Springer, Cham, Switzerland; 2016).
15. Muether, P. S., Hermann, M. M., Koch, K. & Fauser, S. Delay between medical indication to anti-VEGF treatment in age-related macular degeneration can result in a loss of visual acuity. *Graefes Arch. Clin. Exp. Ophthalmol.* **249**, 633–637 (2011).
16. Arias, L. et al. Delay in treating age-related macular degeneration in Spain is associated with progressive vision loss. *Eye* **23**, 326–333 (2009).
17. Karri, S. P. K., Chakraborty, D. & Chatterjee, J. Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration. *Biomed. Opt. Express* **8**, 579–592 (2017).
18. Apostolopoulos, S., Ciller, C., De Zanet, S. I., Wolf, S. & Sznitman, R. RetiNet: automatic AMD identification in OCT volumetric data. Preprint at http://arxiv.org/abs/1610.03628v1 (2016).
19. Farsiu, S. et al. Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography. *Ophthalmology* **121**, 162–172 (2014).
20. Srinivasan, P. P. et al. Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images. *Biomed. Opt. Express* **5**, 3568–3577 (2014).
21. Lee, C. S., Baughman, D. M. & Lee, A. Y. Deep learning is effective for classifying normal versus age-related macular degeneration OCT images. *Ophthalmol. Retin.* **1**, 322–327 (2017).
22. Fang, L. et al. Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search. *Biomed. Opt. Express* **8**, 2732–2744 (2017).
23. Lee, C. S. et al. Deep-learning based, automated segmentation of macular edema in optical coherence tomography. *Biomed. Opt. Express* **8**, 3440–3448 (2017).
24. Lu, D. et al. Retinal fluid segmentation and detection in optical coherence tomography images using fully convolutional neural network. Preprint at http://arxiv.org/abs/1710.04778v1 (2017).
25. Roy, A. G. et al. ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional network. *Biomed. Opt. Express* **8**, 3627–3642 (2017).
26. Castelvecchi, D. Can we open the black box of AI? *Nature* **538**, 20–23 (2016).
27. Schmidt-Erfurth, U. et al. Machine learning to analyze the prognostic value of current imaging biomarkers in neovascular age-related macular degeneration. *Ophthalmol. Retin.* **2**, 24–30 (2018).
28. Schlegl, T. et al. Fully automated detection and quantification of macular fluid in OCT using deep learning. *Ophthalmology* **125**, 549–558 (2018).
29. Keane, P. A. & Sadda, S. R. Predicting visual outcomes for macular disease using optical coherence tomography. *Saudi J. Ophthalmol.* **25**, 145–158 (2011).
30. Schaal, K. B., Rosenfeld, P. J., Gregori, G., Yehoshua, Z. & Feuer, W. J. Anatomic clinical trial endpoints for nonexudative age-related macular degeneration. *Ophthalmology* **123**, 1060–1079 (2016).

31. Schmidt-Erfurth, U. & Waldstein, S. M. A paradigm shift in imaging biomarkers in neovascular age-related macular degeneration. *Prog. Retin. Eye Res.* **50**, 1–24 (2016).
32. Villani, E. et al. Decade-long profile of imaging biomarker use in ophthalmic clinical trials. *Invest. Ophthalmol. Vis. Sci.* **58**, BIO76–BIO81 (2017).
33. Chopra, R., Mulholland, P. J., Dubis, A. M., Anderson, R. S. & Keane, P. A. Human factor and usability testing of a binocular optical coherence tomography system. *Transl. Vis. Sci. Technol.* **6**, 16 (2017).

Author contributions

P.A.K., M.S., J.C., D.H., P.T.K., T.B. and K.A. initiated the project and the collaboration. O.R., J.D.F., B.R.-P. and S.N. developed the network architectures, training and testing setup. P.A.K., J.R.L. and R.C. designed the clinical setup. P.A.K., J.R.L., J.C., R.C., D.A.S., C.E. and A.T. created the dataset and defined clinical labels. J.D.F., B.R.-P., S.N., N.T., S.Bl., H.A., B.O., D.V., G.v.d.D., O.R. and J.C. contributed to the software engineering. J.R.L., S.Bl. and H.A. created the database. P.A.K., J.R.L., D.K., A.K., C.O.H. and R.R. contributed clinical expertise. O.R., P.A.K., J.D.F., J.R.L., B.R.-P., S.N., N.T. and X.G. analysed the data. T.B., S.Bo., J.C., J.H., F.M. and C.M. managed the project. O.R., P.A.K., J.R.L., J.D.F., B.R.-P., G.R. and H.M. wrote the paper. B.L. contributed to the uncertainty estimation.

Competing interests

P.A.K., G.R., H.M. and R.R. are paid contractors of DeepMind. P.A.K. has received speaker fees from Heidelberg Engineering, Topcon, Haag-Streit, Allergan, Novartis and Bayer. P.A.K. has served on advisory boards for Novartis and Bayer, and is an external consultant for DeepMind and Optos. A.T. has served on advisory boards for the following companies: Allergan, Bayer, Genentech, GlaxoSmithKline, Novartis, Roche. C.E. has received speaker fees from Heidelberg Engineering and Haag-Streit UK. P.T.K. has served on advisory boards for Aerie, Allergan, Alcon, Belkin Laser, Novartis and Santen. D.A.S. has received speaker fees from Novartis, Bayer, Allergan, Haag-Streit. The authors have no other competing interests to disclose.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-018-0107-6.

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

Correspondence and requests for materials should be addressed to P.A.K. or O.R.

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

Methods

Ethics and information governance. This work and the collection of data on implied consent received national Research Ethics Committee (REC) approval from the Cambridge East REC and Health Research Authority approval (reference 16/EE/0253); it complies with all relevant ethical regulations. Deidentification was performed in line with the Information Commissioner's Anonymization: managing data protection risk code of practice (https://ico.org.uk/media/1061/anonymisation-code.pdf), and validated by the Moorfields Eye Hospital Information Technology and Information Governance departments, respectively. Only deidentified retrospective data were used for research, without the active involvement of patients.

Visualization of results in clinical practice. To facilitate viewing of the results in routine clinical practice, we display the obtained three-dimensional segmentation maps as two-dimensional thickness maps overlaid on a projection of the raw OCT scan (Fig. 5a). The thickness maps for all tissue types are displayed side-by-side in our interactive OCT viewer (Fig. 5b and Supplementary Video 1). Our system also provides measures for its degree of certainty on both overall referral decision, and each specific retinal disease feature. In most common clinical scenarios, the algorithm will both provide the diagnosis with a high degree of certainty and highlight classical disease features (for example, 'wet' AMD; Supplementary Video 2). This visualization may be particularly useful for difficult and ambiguous cases, such as the diagnosis of choroidal neovascularization formation in cases of chronic central serous retinopathy (Supplementary Videos 5, 7) or in advanced geographic atrophy due to AMD (Supplementary Video 6). Such visualization may also allow clinicians to discard an automated diagnosis or referral suggestion in obvious failure cases, such as when poor image quality leads to erroneous segmentation results (Supplementary Video 8). Furthermore, in a screening context the tissue segmentation map can facilitate quality assurance procedures, whether in normal cases (Supplementary Video 3) or in disease cases (for example, diabetic macular edema in the context of diabetic retinopathy screening, Supplementary Video 4).

Datasets and clinical taxonomy. *Datasets.* Data were selected from a retrospective cohort of all patients who attended Moorfields Eye Hospital NHS Foundation Trust, a world renowned tertiary referral center with 32 clinic sites serving an urban, mixed socioeconomic and ethnicity population centered around London, United Kingdom, between 1 June 2012 and 31 January 2017, who received OCT imaging (Topcon 3D OCT, Topcon; Spectralis, Heidelberg Engineering) as part of their routine clinical care. Conditions with fewer than ten cases, and data from patients who had manually requested that their data should not be shared, were excluded before research began. OCT scan sets containing severe artefacts or marked reductions in signal strength to the point at which retinal interfaces could not be identified were also excluded from the study (Supplementary Fig. 10), as such scans are non-diagnostic and in practice would usually be retaken. Scans to which no diagnostic label could be attached (as described below) were excluded from the present study. For OCT examinations that were labeled as urgent or semi-urgent in the Moorfields OpenEyes electronic health record only scans taken prior to treatment beginning were included; during treatment, resolution of pathology invalidates the database labels. The dataset selection and stratification process is displayed in a CONSORT flow diagram in Supplementary Fig. 11.

Two OCT device types were selected for investigation. 3D OCT-2000 (Topcon, Japan) was selected as device type 1, because of its routine use in the clinical pathway that we studied. For device type 1, a total of 15,877 OCT scans from 7,981 individual patients (mean age 69.5; 3,686 male, 4,294 female, 1 gender unknown) were eligible for inclusion in the work (datasets 3 and 4 in Supplementary Table 3). To create a test set representative of the real-world clinical application, 997 additional patients (mean age 63.1; 443 male, 551 female, 3 gender unknown) presenting to Moorfields with visual disturbance during the retrospective period were selected and only their referral OCT examination was selected for inclusion in the test set (dataset 5 in Supplementary Table 3); a sample size requirement of 553 to detect sensitivity and specificity at 0.05 marginal error and 95% confidence was used to inform the number included. To demonstrate the generalizability of our approach, Spectralis OCT (Heidelberg Engineering) was chosen as 'device type 2'. For generalizability experiments, a second test set of clinical OCT scans from 116 patients (mean age 58.2; 59 male, 57 female) presenting in the same manner were selected using the same methodology and selection criteria (dataset 11 in Supplementary Table 3). Examples of differences between the two devices types are shown in Supplementary Fig. 9. Supplementary Table 8 shows a breakdown of patients and triage categories in the datasets.

Clinical taxonomy. OCT examinations were mapped from individual diagnoses and treatment information to specific triage decisions (urgent referral, semi-urgent referral, routine referral and observation only) to a medical retina clinic setting (Supplementary Table 1). Where possible, the presence or absence of additional pathologies was added as a label (Supplementary Table 5). The dataset represents the full variety of medical retina patients presenting and receiving treatment at Moorfields Eye Hospital. Although the exact mapping was chosen to be relevant to the triage decisions at Moorfields Eye Hospital where the research work took place, the framework is generalizable to other systems at centers with different

triage requirements (for example, optometrists working in a high-street clinic setting or ophthalmologists without subspecialty retinal expertise). Scans meeting the exclusion criteria were removed from the database before splitting the data into training, validation and test sets. Supplementary Figure 12 provides an example of variation within the 'urgent referral' label class.

Clinical labeling. Clinical labels for the 14,884 scans in dataset 3 in Supplementary Table 3 were assigned through an automated notes search with trained ophthalmologist and optometrist review of the OCT scans. The presence or absence of choroidal neovascularization, referable macular edema, normal and other pathologies visible on the OCT scan were recorded. In addition, patients with choroidal neovascularization or macular edema confirmed through treatment were labeled directly from the Moorfields OpenEyes electronic health record. A validation subset of 993 scans (993 patients) was graded separately by three junior graders (ophthalmologists specializing in medical retina) with disagreement in clinical labels arbitrated by a senior retinal specialist with over 10 years of experience and image reading center certification for OCT segmentation (dataset 4 in Supplementary Table 3). The test set was further verified by full review of the notes with access to follow up data with both junior and senior grader review. Junior and senior graders were separate to those participating in the evaluation of expert performance.

Manual segmentation. A subset of 1,101 scans from device type 1 and a set of 264 scans from device type 2 were manually segmented using the segmentation editor plugin for ImageJ (Fiji)[34] (datasets 1, 2, 9 and 10 in Supplementary Table 3). The segmentation labels were chosen to distinguish all relevant diagnoses for the referral decision, as well as potential artefacts that may affect the diagnostic quality of the whole or part of the scan. In particular, the current state of art does not differentiate between the three different types of pigment epithelial detachment, or segment out areas of fibrosis scarring or blood as hyperreflective material[27,28]. Anatomical delineations and nomenclature are consistent with standard grading criteria for the evaluation of OCT[35–37]. The segmentation examples were selected and segmented by ophthalmologists specializing in medical retina as representative cases for pathological features. These were reviewed and edited by a senior ophthalmologist with over 10 years of experience and image reading center certification for OCT segmentation. Per OCT, 3–5 slices were chosen for segmentation, which best represented the pathological features (Supplementary Tables 2, 9 and Supplementary Fig. 13).

Evaluating the expert performance. To evaluate expert performance on the test set, eight clinical experts were recruited for an evaluation study. Participants included four consultant ophthalmologists at Moorfields Eye Hospital with fellowship-level subspecialty training in medical retinal disease and extensive clinical experience (21, 21, 12.5 and 11.5 years of experience) and four optometrists at Moorfields Eye Hospital with specialist training in OCT interpretation and retinal diseases (15, 9, 6 and 2.5 years of experience). These are referred to as retinal specialists 1–4 and optometrists 1–4 in the rest of the paper (Supplementary Table 10). Each expert was instructed to provide a triage decision (Supplementary Table 1) and to record the presence or absence of defined pathological features (Supplementary Table 5).

To assess the performance in a realistic clinical environment, all scans were read in a random order twice with at least a week between readings. During the initial review, only the OCT scan was presented (dataset 7 in Supplementary Table 3). During the second review, participants were presented with all the information available at the time of triage: OCT and fundus scans, age, gender, ethnicity and where available information on visual acuity and a short clinical vignette (dataset 8 in Supplementary Table 3). The model only received the OCT scan.

To assess the difference between the test set for device type 1 and device type 2, five clinical experts were recruited for a further evaluation study (dataset 12 in Supplementary Table 3). Participants were five consultant ophthalmologists at Moorfields Eye Hospital with fellowship-level subspecialty training in medical retinal disease (21, 21, 12.5, 11.5 and 11 years of experience). Four were participants in the device type 1 evaluation study, while the other was a new participant for this study and is referred to as retina specialist five.

Network architectures and training protocol. *Segmentation network.* The first stage of our framework consists of a segmentation network that takes as input part of the OCT scan, and outputs a part of a segmentation map. That is, it predicts for each voxel one tissue type out of the 15 classes described in Supplementary Table 2. At training time, the input of the network consists of 9 contiguous slices of an OCT, and the goal of the network is to segment the central slice. The input is therefore a $448 \times 512 \times 9$ voxels image, and the output is an estimated probability over the 15 classes, for each of the $448 \times 512 \times 1$ output voxels. None of the convolutions made across the slices (z dimension) adds padding to its input. As a result, we can exploit shared computations at inference time to predict any number of contiguous slices in parallel, which was only limited by the memory capacity of the system.

The structure of the segmentation convolutional neural network model is shown in Supplementary Fig. 14. It uses a three-dimensional U-Net architecture[14], consisting of an analysis (downwards) path, a synthesis (upwards) path, and shortcut connections between blocks of the same level and different paths.

We applied four variations over it. First, we used $3 \times 3 \times 1$ convolutions with padding and $1 \times 1 \times 3$ convolutions without padding instead of $3 \times 3 \times 3$ convolutions without padding. Second, downsampling and upsampling operations were carried out through parameter-free bilinear interpolation, replacing max-pooling and up-convolution. Third, we introduced one extra residual connection within each block of layers, so that the output of each block consists of the sum of the features of the last layer, and the first layer of the block in which the features dimensions match. Finally, the middle block of layers between the analysis and synthesis paths is composed of a sequence of fully connected layers. The first variation allows us to control the receptive field for z separately and is furthermore less computationally intensive. The second and third variation aimed at improving the gradients flow throughout the network, which makes the training process easier. The last variation extends the receptive field such that each pixel in the output effectively has the whole input contained within its receptive field.

We used per-voxel cross entropy as the loss function, with 0.1 label-smoothing regularization[38]. We have neither used dropout nor weight decay as regularization means, as preliminary experiments showed that this did not improve the performance. We trained the model in TensorFlow[39] with the Adam optimizer[40] for 160,000 iterations on 8 graphics processing units (GPUs) with dataset 1 in Supplementary Table 3. The initial learning rate was 0.0001 and set to 0.0001/2 after 10% of the total iterations, 0.0001/4 after 20%, 0.0001/8 after 50%, 0.0001/64 after 70%, 0.0001/256 after 90% and finally 0.0001/512 for the final 5% of training. All decisions and hyperparameters above were selected on the basis of their performance on a validation set (dataset 2 in Supplementary Table 3).

To improve the generalization abilities of our model, we augmented the data by applying affine and elastic transformations jointly over the inputs and ground-truth segmentations[13,14]. Intensity transformations over the inputs were also applied.

Our segmentation network for device type 2, which is shown in Supplementary Fig. 15, is trained on scans from both devices (datasets 1 and 9 in Supplementary Table 3) with the aim of leveraging the large number of labeled instances for device type 1. It has three changes compared to the architecture for device type 1. First, we subsample the input from device type 1 (128 slices) to match the resolution of device type 2 (49 slices) and apply slight padding in height to the scans of device type 2 to give them of the same shape in height and width as the scans of device type 1. Second, the input first goes through one of two 'device adaptation branches', depending on the device type of the input scan. The architecture of this branch consists of three convolutions with padding, with one residual connection as in the other blocks, and is identical for both device types (see Supplementary Fig. 15). The network can then simply learn to compensate for the changes between device types early on and map them to a common representation. Lastly, the number of feature maps on the first level of the analysis path is halved from 32 to 16 such that the overall architecture still has fewer parameters than the architecture for device type 1. During training, the network was presented with a ratio of 2.5: 1 for training samples from device type 2:device type 1. All decisions and hyperparameters above were selected on the basis of their performance on a validation set (datasets 2 and 10 in Supplementary Table 3).

Classification network. The classification network learned to map a segmentation map to the four referral decisions and the ten additional diagnoses (see Supplementary Fig. 16). For device type 1, it takes as input a $300 \times 350 \times 43$ subsampling of the original $448 \times 512 \times 128$ segmentation map created by the segmentation network described above. The output is a 14-component vector. For device type 2, for which the scans originally were $448 \times 512 \times 49$, we first upscaled the segmentation map to the same resolution as for device type 1 and then proceed identically as for device type 1. The architecture uses a three-dimensional version of the dense blocks described previously[41] using $3 \times 3 \times 1$ and $1 \times 1 \times 3$ convolutions. The details of its structure are shown in Supplementary Fig. 16. We found using dense convolution blocks to be critical for training classification networks on large three-dimensional volumes. The inputs are one-hot encoded and augmented by random three-dimensional affine and elastic transformations[14]. The loss was the sum of the softmax cross entropy loss for the first four components (multi-class referral decision) and the sigmoid cross entropy losses for the remaining ten components (additional diagnoses labels). We also used a small amount (0.05) of label-smoothing regularization[38] and added some (1×10^{-5}) weight decay. We trained the model in TensorFlow[39] with the Adam optimiser[40] for 160,000 iterations of batch size 8 spread across 8 GPUs with 1 sample per GPU with dataset 3 in Supplementary Table 3. The initial learning rate was 0.02 and set to 0.02/2 after 10% of the total iterations, 0.02/4 after 20%, 0.02/8 after 50%, 0.02/64 after 70%, 0.02/256 after 90% and finally 0.02/512 for the final 5% of training. All decisions and hyperparameters described above were selected on the basis of their performance on a validation set (dataset 4 in Supplementary Table 3).

Ensembling. For both of these networks we trained five instances. We trained the same network with a different order of the inputs and different random weight initializations[42]. Previously published experiments[42] suggest that five instances are sufficient in most settings, so we also used this number. For our experiments, we applied the five instances of our segmentation model to the input scan resulting in five segmentation maps. The five instances of our classification model were then

applied to each of the segmentation maps, resulting in a total of 25 classification outputs per scan, as illustrated in Supplementary Fig. 1. The results reported are obtained after averaging the probabilities estimated by these models.

Optimizing the ensemble output for sensitivity, specificity and penalty scores. For different applications, the preferred compromise between a high hit rate (sensitivity) and a low false alarm rate ($1 - $ specificity) can be different. For the binary diagnosis decisions, we computed an optimal rescaling factor a for the pseudo-probabilities, such that a 50% threshold achieves maximal (sensitivity + specificity)/2 on the validation set (dataset 4 in Supplementary Table 3). The rescaling was done by $p = aq/(aq + (1-a)(1-q))$, where q denotes the ensemble output and p the reweighted probability. We used (sensitivity + specificity)/2 instead of the total accuracy to avoid the bias due to the low number of patients with a positive condition in the validation set (and in the test set). For a balanced set with equal numbers of positive and negative samples this term is exactly the accuracy.

For the four-way referral decision (where the highest probability wins), we optimized four scaling factors using the validation set to reduce the overall cost specified by the misclassification penalty matrix (Supplementary Fig. 6). A first set of factors was optimized for a balance between high accuracy and low penalty points (referred to as "our model (1)" in Supplementary Figure 6), a second set of factors was optimized for penalty cost only (referred to as "our model (2)" in Supplementary Figure 6). The cost matrix for the balanced performance was computed by averaging the normalized cost matrix for accuracy (a matrix with 0 in the diagonal elements and 1 in the off-diagonal elements) and the normalized penalty cost matrix. Normalization was performed by dividing the matrix by the sum of all elements. The optimization of the four factors was done with the Adam optimiser using a softmax layer and a weighted cross-entropy loss layer.

End-to-end classification network. The network architecture for the end-to-end classification experiments was identical to the architecture of the classification network in the two-stage approach (see 'Classification network' and Supplementary Fig. 16) with a small adaption. To roughly obtain the same number of parameters, we added a dense layer (two convolutions with seven channels output each) that translates the single-channel raw OCT to a 14-channel feature map. All selected hyperparameters and augmentation strategies were identical to the original classification network. We trained five network instances on the training set with 14,884 raw OCT scans from device type 1 (dataset 13 in Supplementary Table 3). Each network instance was initialized with different random weights and was presented with the training images in a different order. After training, we also computed an optimal reweighting on the validation set (as we did for the two-stage model) and tested the ensemble on the test set.

Statistical analysis. *Significant differences using a two-sided exact binomial test.* The comparison of our model's performance to the expert's performance is based on the assumption that our model and the expert have an unknown but constant performance. That is, every inspected eye scan is correctly diagnosed by our model with the probability p_{mod}, and correctly diagnosed by the expert with probability p_{exp}. For N eye scans the number of correct diagnoses k is therefore binomially distributed with $\Pr(k) = B(k|p,N)$. If our model achieves k_{mod} correct diagnoses and the expert achieves k_{exp} correct diagnoses, the probability that the true performance of our model p_{mod} is higher than the true performance of the expert p_{exp} is

$$\Pr(p_{mod} > p_{exp} \mid k_{mod}, k_{exp}, N) =$$

$$\frac{\int_0^1 B(k_{mod} \mid p_1, N) \int_0^{p_1} B(k_{exp} \mid p_2, N) \, dp_2 \, dp_1}{\int_0^1 B(k_{mod} \mid p, N) \, dp \int_0^1 B(k_{exp} \mid p, N) \, dp}$$

The probability for a lower performance, that is $\Pr(p_{mod} < p_{exp} | k_{mod}, k_{exp}, N)$ is derived analogously. For all comparisons, a confidence level of 95% was used. The formula was numerically integrated using in-house code.

Further details on the methods are described in a published protocol describing the DeepMind collaboration with Moorfields Eye Hospital[43].

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Code availability. The code base for the deep-learning framework makes use of proprietary components and we are unable to publicly release the full code base. However, all experiments and implementation details are described in sufficient detail in the Methods and in the Supplementary Figs. to enable independent replication with non-proprietary libraries. The three-dimensional augmentation code (using the caffe framework) is available as part of the three-dimensional U-net source code at https://lmb.informatik.uni-freiburg.de/resources/opensource/unet.en.html. Additionally, although we are unable to make all the Google proprietary components available, we are in the process of making the augmentation operations for TensorFlow available in the official TensorFlow code.

Data availability. The clinical data used for the training, validation and test sets were collected at Moorfields Eye Hospital and transferred to the DeepMind data center in the UK in deidentified format. Data were used with both local and national permissions. They are not publicly available and restrictions apply to their use. The data, or a test subset, may be available from Moorfields Eye Hospital NHS Foundation Trust subject to local and national ethical approvals.

References

34. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
35. Keane, P. A. et al. Evaluation of age-related macular degeneration with optical coherence tomography. *Surv. Ophthalmol.* **57**, 389–414 (2012).
36. Folgar, F. A. et al. Comparison of optical coherence tomography assessments in the comparison of age-related macular degeneration treatments trials. *Ophthalmology* **121**, 1956–1965 (2014).
37. Duker, J. S., Waheed, N. K. & Goldman, D. *Handbook of Retinal OCT: Optical Coherence Tomography E-Book* (Elsevier Health Sciences, Oxford, UK; 2013).
38. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 2818–2826 (2016).
39. Abadi, M. et al. TensorFlow: large-scale machine learning on heterogeneous systems. Preprint at https://arxiv.org/abs/1603.04467 (2016).
40. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. in Proceedings of the 3rd International Conference on Learning Representations (ICLR). Preprint at http://arxiv.org/abs/1412.6980 (2015).
41. Huang, G., Liu, Z., Weinberger, K. Q. & van der Maaten, L. Densely connected convolutional networks. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 2261–2269 (2017).
42. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Adv. Neural Inf. Process. Syst.* 6405–6416 (2017).
43. De Fauw, J. et al. Automated analysis of retinal imaging using machine learning techniques for computer vision. *F1000Res* **5**, 1573 (2016).

naturereseach

Corresponding author(s):   Olaf Ronneberger

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

▶ Experimental design

1. Sample size

Describe how sample size was determined.

A sample size requirement of 553 to detect sensitivity and specificity at 0.05 marginal error and 95% confidence was used to inform the number included in the test set. A sample of 997 patients were selected to be part of the gold standard test set against which the human experts and the model were compared.

The total of 15,877 TopCon 3D OCT 2000 scans from 7981 individual patients were eligible for inclusion in the work. An additional 268 Heidelberg Spectralis scans were selected in order to conduct generalisability experiments. The total sample size for training and validation sets was informed by the existing literature and by DeepMind's previous work in the field of machine learning (Mnih et al., 2015; Silver et al., 2016). Today's most powerful deep neural networks can have millions or billions of parameters, so large amounts of data are needed to automatically infer those parameters during learning. Most problems in the medical domain are highly complex as they arise as an interplay of many clinical, demographic, behavioural and environmental factors that are correlated in non-trivial ways. This is even more true for state-of-the art deep learning methodologies that are expected to give the best results (Szegedy et al., 2014).

2. Data exclusions

Describe any data exclusions.

OCT image sets with no diagnostic labels, those containing severe artefacts, or significant reductions in signal strength to the point where retinal interfaces could not be identified were excluded from the present study. Conditions with fewer than ten cases, and data from patients who had manually requested that their data should not be shared, were excluded before research began. For the test set patients who had previously been treated in clinic by the evaluation study participants were excluded from the test set. For more detail please refer to the manuscript methods section.

3. Replication

Describe whether the experimental findings were reliably reproduced.

All 997 patients in the test set for the first device type were randomly selected and were not correlated in any way. The experiments can be interpreted as 997 replicas of a single patient diagnosis. Without retraining the classification network in our framework performance was reproduced on a new test dataset from a second device type of 116 OCT scans. The performance in each case is as follows: Device Type 1 error rate: 55 out of 997 = 5.5%; Device Type 2 error rate: 4 out of 116 = 3.4%.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Samples meeting the inclusion criteria were randomly allocated to training or validation sets. A separate group of patients were randomly selected before creation of the training and validation datasets as an independent test set which was kept separate during model development. Randomisation was on individual patients rather than OCT images: where there were multiple scans for a single patient these were allocated to only one of training, validation or test. For more detail please refer to the manuscript methods section.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Participants in the clinical evaluation of the models were blinded to the ground truth and were not involved in dataset collection; patients who had previously been treated in clinic by the participants were excluded from the test set.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The underline exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

▶ Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

The networks used the TensorFlow library with custom extensions (see methods section). Analysis was performed with custom code written in Python.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

▶ Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

The clinical data used for the training, validation and test sets were collected at Moorfields Eye Hospital and transferred to DeepMind data centre in the UK in de-identified format. Data were used with both local and national permissions. They are not publicly available and restrictions apply to their use. The data, or a test subset, may be available from Moorfields Eye Hospital subject to local and national ethical approvals.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

No eukaryotic cell lines were used.

b. Describe the method of cell line authentication used.

No eukaryotic cell lines were used.

c. Report whether the cell lines were tested for mycoplasma contamination.

No eukaryotic cell lines were used.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

No eukaryotic cell lines were used.

▶ Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

> No animals were used in the study.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

> Data were selected from a retrospective cohort of all patients attending Moorfields Eye Hospital NHS Foundation Trust, a world renowned tertiary referral centre with multiple clinic sites serving an urban, mixed socioeconomic and ethnicity population centred around London, U.K., between 01/06/2012 and 31/01/2017, who had OCT imaging (Topcon 3D OCT, Topcon, Japan; Spectralis, Heidelberg, Germany) as part of their routine clinical care. For more details please refer to the manuscript methods section.
>
> Two OCT device types were selected for investigation. 3D OCT-2000 (Topcon, Japan) was selected as "device type 1" due to its routine use in the clinical pathway we studied. For device type 1, a total of 15,877 OCT scans from 7981 individual patients (mean age 69.5; 3686 male, 4294 female, 1 gender unknown) were eligible for inclusion in the work (Datasets #3 + #4 in Supplementary Table 3). To create a test set representative of the real-world clinical application, 997 additional patients (mean age 63.1; 443 male, 551 female, 3 gender unknown) presenting to Moorfields with visual disturbance during the retrospective period were selected and only their referral OCT examination was selected for inclusion in the test set (Dataset #5 in Supplementary Table 3); a sample size requirement of 553 to detect sensitivity and specificity at 0.05 marginal error and 95% confidence was used to inform the number included. To demonstrate the generalizability of our approach, Spectralis OCT (Heidelberg Engineering, Germany) was chosen as "device type 2". For generalisability experiments, a second test set of clinical OCT scans from 116 patients (mean age 58.2; 59 male, 57 female) presenting in the same manner were selected using the same methodology and selection criteria (Dataset #11 in Supplementary Table 3). Examples of differences between the two devices types are shown in Supplementary Fig. 9. Supplementary Table 8 shows a breakdown of patients and triage categories in the datasets.