TITLE:

Classifying the severity of diabetic macular oedema from optical coherence tomography scans using deep learning: a feasibility study.

AUTHORS:

Cathal Breathnach (1), Fiona Harney (1,2), Deirdre Townley (2), Rachel Hickey (1), Andrew Simpkin (3), Derek O'Keeffe (1,4).

(1) School of Medicine, College of Medicine Nursing and Health Sciences, University of Galway, Galway.
(2) Department of Ophthalmology, University Hospital Galway, Galway.
(3) School of Mathematical and Statistical Sciences, University of Galway, Galway, Ireland.
(4) Department of Endocrinology, University Hospital Galway, Galway.

ABSTRACT:

Background: Diabetic macular oedema (DME) is a vision-threatening complication of diabetes mellitus. It is reliably detected using optical coherence tomography (OCT). This work evaluates a deep learning system (DLS) for the automated detection and classification of DME severity from OCT images.

Methods: Anonymised OCT images were retrospectively obtained from 950 patients at University Hospital Galway, Ireland. Images were graded by a consultant ophthalmologist to classify the level of DME on a novel scale (Normal, DME not affecting the foveal contour and DME affecting the foveal contour), excluding other pathologies. A DLS was trained using cross-validation, then evaluated on a test dataset and an external dataset. The test set was graded by a second ophthalmologist.

Results: In detecting the presence of DME, the DLS achieved a mean area under the receiver operating characteristic curve (AUC) of 0.98 on cross-validation. AUCs of 0.94 (95% CI 0.90-0.98) and 0.91 (0.86-0.95) were achieved on evaluation of DME detection for the test dataset when graded by the first and second ophthalmologist respectively. An AUC of 0.94 (0.92-0.96) was achieved on evaluation with the external dataset. When detecting the DME severity, AUCs of 0.97, 0.87 and 0.98 were achieved per class on cross validation. For the test dataset, AUCs of 0.93, 0.83 and 0.97 were achieved when graded by the first ophthalmologist and AUCs of 0.89, 0.75 and 0.96 were achieved when graded by the second ophthalmologist.

Conclusion: This study suggests promising results for the use of deep learning in the classification of severity of DME.

INTRODUCTION

Diabetic macular oedema (DME) is the accumulation of fluid in the macula that can occur at any stage of diabetic retinopathy (DR).[1] DME is associated with more severe DR, and can result in visual loss.[2] Early detection facilitates treatment, including systemic control of diabetes alongside treatments such as laser photocoagulation and intraocular anti-vascular endothelial growth factor injections.[3] Many DR screening programmes use retinal photographs for DME diagnosis, but this relies on surrogate markers, which is associated with high false positive rates, leading to unnecessary referrals.[3–5] In comparison, optical coherence tomography (OCT) provides a detailed structural view of the retina and is the gold standard for DME evaluation.[4,6] OCT is quick, non-invasive and has been recommended for use in screening programmes to reduce unneeded referrals and cost.[4] Certain screening programmes already routinely use OCT for example in Denmark,[7] and in Ireland, where it is used for higher risk patients and has been predicted to reduce referrals.[8] However, screening processes are often manual, and OCT scans require expert interpretation.[2,9]

Deep learning (DL) is a form of artificial intelligence (AI) that can be used to analyse OCT scans.[9,10] A convolutional neural network (CNN) is a class of DL algorithm that is commonly used for medical image classification,[11] and have been shown to reliably detect DME from OCT scans.[12] Several approaches for the implementation of DL algorithms for OCT analysis have previously been tested.[13–24] Some studies have analysed two-dimensional images from OCT images (B-scans) for the detection of DME and other retinal pathologies.[13–15] Other studies have used a transfer learning approach (which uses an existing model trained to recognise general images and then optimises it for a supplied dataset) which can offer an advantage when very large datasets are not available.[11] This approach has been tested for the detection of DME and other retinal pathologies such as choroidal neovascularization and drusen.[11,16,17] Several studies have specifically analysed three-dimensional OCT volumes to detect DME alongside age-related macular degeneration.[18,19] Other studies have taken an approach for full segmentation of the retina to diagnose disease, using an additional segmentation step to allow the detection of various pathologies or to segment retinal fluid.[20–22] A notable study was developed by De Fauw et al., using an initial segmentation approach to reliably predict a referral level.[23] However, while each study reports favourable performance for the detection of DME and other retinal pathologies, there are limitations. For example, the segmentation approach requires labour intensive annotation and approaches such as transfer learning can require considerable input to fine-tune the algorithm.

Furthermore, few studies including those outlined thus far consider the severity of the DME present, only detecting whether it is present or absent. Various terms have been used to describe DME, including clinically-significant diabetic macular oedema, which can be determined by the presence of specific characteristics.[25] The central area of the fovea is of most interest, and DME within 500μm on either side of the foveal centre is referred to as centrally-involved DME (CI-DME).[2] The International Council of Ophthalmology (ICO) recommends that the presence of CI-DME should be used in conjunction with visual acuity to guide treatment

decisions.[2] A study by Tang et al. considered this difference when testing a DL approach for DME detection and classification.[24] This study achieved favourable performance and was tested using both two-dimensional images and three-dimensional scan data. There are also specific biomarkers detectable using OCT scans that can be used to define the DME severity,[26] and a study by Xu et al. used DL to analyse for the presence of such features, although this study required segmentation.[27]

With the exception of the study completed by Tang et al.,[24] no other studies have classified the severity of DME from OCT scans using a standard approach and it is clear this area is under-explored. Our study will implement a simple and lightweight approach to this problem, using a novel scale for DME severity. Furthermore, we are not aware of any study that has previously tested DL for the detection of diabetic eye disease in an Irish population.

MATERIALS & METHODS

Image Collection & Datasets

Patients were randomly selected from all patients with diabetes currently being treated in University Hospital Galway under the Irish national DR screening programme to create the training and validation dataset. Images were taken from the foveal level of one scan per patient, taken with a Heidelberg Spectralis OCT machine. A square image, centred around the fovea, was used from both eyes when available. As a retrospective study, there was no patient follow-up.

A second test dataset was created using scan images from every patient in the diabetic registry who attended for an ophthalmology appointment over a two-week period in July 2022. No patients from the training and validation dataset were included in the test dataset. Finally, an externally graded dataset of 250 images, assessed by multiple graders, was also used to test the algorithm.[28] Only DME detection was tested using this external set as these images were not graded using the scheme proposed in this study.

Grading Criteria

This study aimed to distinguish between normal OCT images and those with signs of DME, but it also classified the level of DME present on a novel scale. This scale distinguishes DME severity based on the presence of disruption to the normal foveal contour, as outlined in Table 1. The images in Figure 1 provide examples of images as per this scheme.

The central area is taken is taken as 500μm on either side of the foveal centre in this study, in keeping with ICO guidelines for the treatment of DME.[2] This is the area of most significance as the central area may be the thickest part of the retina with DME.[25] This is a similar approach to that taken for the automated detection of CI-DME,[24] but DME that disrupts the foveal contour represents more severe DME.

DME Classification	Requirements
DME-0	No signs of DME.
DME-1	Signs of DME present with no change to the foveal contour in the central area*.
DME-2	Signs of DME present with a change to the foveal contour in the central area*.

Table 1: DME classification requirements. *The central area is defined as 500µm on either side of the foveal centre.

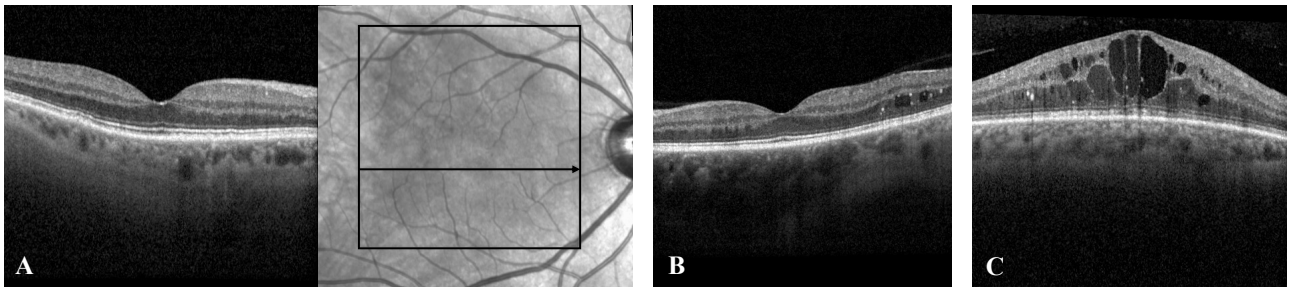

Figure 1: OCT image examples as per the grading scheme used in this study, showing an example of DME-0 with corresponding level on a retinal photograph (A), an example of DME-1 showing DME with no change to the foveal contour (B) and an example of DME-2 showing DME-2 and loss of the normal foveal contour (C).

Exclusion Criteria

Exclusion criteria were used to exclude non-relevant images for the scope of this study. The study excluded (i) poor quality scans, where the retinal layers at the level of the fovea are not discernible, (ii) images with signs of other retinal pathologies, an example of which is wet age-related macular degeneration, (iii) loss of the foveal contour caused by another retinal pathology, for example due to an epiretinal membrane and (iv) images that did not represent the foveal level.

Grading Process

The training and validation dataset was graded by an ophthalmology consultant (first ophthalmologist) as per the grading scheme outlined. This ophthalmologist also determined which images met the exclusion criteria for the training and validation data.

The test dataset was graded by the first ophthalmologist and by a second ophthalmology consultant (second ophthalmologist) for comparison. The graders were both aware of the exclusion criteria in this case and there was agreement on which images to exclude. Both graders discussed the grading criteria to reach a consensus but then graded the images independently. Cohen's k [29] was also calculated as a measure of agreement. The algorithm trained with the training and validation data generated predictions for the test dataset and were compared independently to the grades produced by the two ophthalmologists. Both consultant ophthalmologists involved in this study had more than 10 years of experience in reading OCT scans and treating diabetic eye disease. No clinical details were provided for image grading.

Algorithm Overview

A CNN was designed and implemented using the Python programming language (version 3.8.13) with the Keras (version 2.9.0) and Tensorflow frameworks (version 2.9.2). The training and validation stage used stratified 5-fold cross validation. The structure of the model is outlined in Supplemental Table 1 and is approximately similar to CNNs used in other DL studies for OCT image analysis.[13] The overall structure and sizes used were determined experimentally to optimise performance within the scope of this study. Two near-identical models were trained in this study with only a differing output layer. The first model for DME detection, used the sigmoid activation for two-class classification with a threshold of 0.5 and the second model used the softmax activation function for three-class classification of the level of DME. Each convolution layer used zero padding with a relu activation function. The model was trained for 35 epochs in each setting with a learning rate of 0.001, using the Adam optimiser.[30] Data augmentation was used as has been implemented in similar studies,[24] using the Keras library to reduce overfitting. This involved random flipping, 10% random rotation and 10% random zoom applied to the data before the rescaling step.

The performance of the algorithm was measured by reporting accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and area under the receiver operating characteristic curve (AUC). Confidence intervals for the performance indices were calculated in the test data using the DeLong method for AUC[31] (*pROC* R package, version 1.18.5) and the Wilson method otherwise [32] using the *epiR* package (version 2.0.77).

RESULTS

Population

Within the available timeframe, a total of 1878 images were included from 950 distinct patients. The training and validation dataset was comprised of 1598 images from 807 patients, and after the exclusion criteria were applied, there were 1501 images available from 793 patients. The test dataset was comprised of 280 images from 143 patients, and after the exclusion criteria were applied, there were 239 images included from 132 patients. The class breakdown is similar between the training and test datasets and is presented with other dataset characteristics in Table 2, graded as per the first ophthalmologist. The test dataset was graded by both the first and second ophthalmologist, and Cohen's κ was calculated as 0.82 for DME detection suggesting near-perfect agreement on DME detection and 0.78 for DME classification suggesting substantial agreement.

Detection of DME

The results for the detection of DME are presented in Table 3. This includes the results from the cross-validation training and then compares the trained algorithm results to the gradings from the first ophthalmologist and the second ophthalmologist. The results are also included when the algorithm is tested with the external dataset.

The confusion matrices and the AUC curves for the grading completed by the first and second ophthalmologist and for the external dataset are included as supplemental Figures 1 and 2.

Characteristic	Training Dataset		Test Dataset	
	n	%	n	%
Included Patients	793		132	
Included Images	1501		239	
Patient Demographics				
Mean age (years)	65.00		67.83	
Age SD (years)	14.0		13.84	
Male Proportion	486	61.4	82	62.1
DME Grade (Images)				
DME-0	943	62.8	134	56.1
DME-1	200	13.3	49	20.5
DME-2	358	23.9	56	23.4

Table 2: Characteristics of the training and test data collected for all non-excluded images. Percentage values expressed relative to the training or test datasets. Graded as per the first ophthalmologist.

Classification of DME

The mean accuracy overall for the detection of the DME type when evaluated for the 5-fold cross-validation was 88.94% (SD = 3.65%). When the algorithm for DME classification was evaluated using the test dataset graded by the first ophthalmologist, the accuracy of the algorithm was 86.61%. When compared to the second ophthalmologist, the accuracy of the algorithm was 80.03%. The AUC values per class are presented in Table 4 with other metrics for the cross-validation and test data. The confusion matrices and the AUC curves comparing to the DME classification algorithm are included as supplemental Figures 3 and 4.

	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	AUC
Training Data: Cross-Validation	94.4 (1.3)	90.8 (6.5)	96.2 (3.1)	94.0 (5.3)	95.0 (3.4)	0.98 (0.01)
Test Data: First Ophthalmologist	92.5 (88.8-95.2)	91.4 (84.5-95.4)	93.3 (87.7-96.4)	91.4 (84.5-95.4)	93.3 (87.7-96.4)	0.94 (0.90-0.98)
Test Data: Second Ophthalmologist	87.9 (83.2-91.4)	83.3 (75.4-89.1)	92.0 (85.9-95.6)	90.5 (83.4-94.8)	85.8 (78.9-90.7)	0.91 (0.86-0.95)
External Dataset	89.40 (86.4-91.8)	92.0 (88.0-94.8)	86.8 (82.0-90.4)	87.5 (82.9-90.9)	91.6 (87.3-94.5)	0.94 (0.92-0.96)

AUC, area under the receiver operating characteristic curve; NPV, negative predictive value; PPV, positive predictive value.

Table 3: Results for DME detection using the dataset gradings. Values reported as mean (SD) for the cross-validation data The trained algorithm is then compared to the ophthalmologist gradings for the test dataset and to the external dataset, presented with 95% confidence intervals.

	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	AUC
Cross Validation						
DME-0	94.2 (3.6)	96.0 (6.0)	90.79 (1.3)	94.7 (0.5)	94.2 (8.8)	0.97 (0.01)
DME-1	89.6 (3.7)	55.5 (1.7)	94.7 (5.2)	67.4 (1.5)	93.4 (2.1)	0.87 (0.05)
DME-2	96.1 (2.1)	88.4 (7.3)	95.8 (3.1)	88.0 (7.3)	96.5 (2.1)	0.98 (0.05)
Test Data						
1st Ophthalmologist						
DME-0	90.8 (86.5-93.8)	97.8 (93.6-99.3)	81.9 (73.5-88.1)	87.3 (81.1-91.7)	96.6 (90.6-98.9)	0.93 (0.90-0.97)
DME-1	88.7 (84.1-92.1)	53.1 (39.4-66.3)	97.9 (94.7-99.2)	86.7 (70.3-94.7)	89.0 (84.0-92.6)	0.83 (0.75-0.91)
DME-2	93.7 (89.9-96.2)	89.3 (78.5-95.0)	95.1 (90.9-97.4)	84.7 (73.5-91.8)	96.7 (92.9-98.5)	0.97 (0.95-0.99)
2nd Ophthalmologist						
DME-0	87.0 (82.2-90.7)	97.6 (93.2-99.2)	75.4 (66.8-82.4)	81.3 (74.3-86.8)	96.6 (90.6-98.9)	0.89 (0.84-0.94)
DME-1	82.0 (76.6-86.4)	40.3 (29.4-52.3)	98.3 (95.0-99.4)	90.0 (74.4-96.5)	80.9 (75.0-85.6)	0.75 (0.67-0.83)
DME-2	91.6 (87.4-94.5)	91.5 (80.1-96.6)	91.7 (87.0-94.8)	72.9 (60.4-82.6)	97.8 (94.4-99.1)	0.96 (0.94-0.99)

AUC, area under the receiver operating characteristic curve; NPV, negative predictive value; PPV, positive predictive value.

Table 4: Results for the DME classification using the dataset grading. Values reported as mean (SD) for the cross-validation data. The trained algorithm is then compared using the test dataset as graded by both the first and second ophthalmologist with values presented with 95% confidence intervals.

DISCUSSION

The algorithm in this study demonstrates excellent performance for the detection of DME on the cross-validated training and similar performance on the test data when compared to the gradings by the first ophthalmologist, achieving an accuracy of 92.5% and an AUC of 0.94. When compared to the second ophthalmologist, the accuracy of the algorithm is slightly reduced at 87.9% with an AUC of 0.91 although the specificity and PPV is similar in both cases. Given that the first ophthalmologist produced the gradings for the training dataset, this is likely a source of the slight difference in accuracy. The similar specificity however suggests the second ophthalmologist may have a slightly lower threshold for diagnosing DME.

The DME detection algorithm also showed favourable performance when compared to the external dataset, with an AUC of 0.94, matching the test dataset performance for the first ophthalmologist. The previous study using this dataset did achieve a higher AUC of 0.99 for binary classification using this as a test dataset, but not as an external dataset and multiple images from the same patient were used in this study.[11] The images used in this

dataset also covered a wider section of the retina and could have been interpreted differently when re-sized for input in this study. There can also be slight ethnic difference between populations and has been suggested this could cause slight differences in algorithm performance when using OCT scans.[24] However, while these slight differences might contribute to variance between algorithms when tested on different populations, this study overall has demonstrated reliable performance for the detection of DME in an Irish population, where the use of DL for OCT image analysis has not previously been tested.

The DME classification algorithm in this study has also shown promising performance. The cross-validation training shows high performance, particularly for the detection of the DME-0 and DME-2 class. When compared to the first ophthalmologist using the test data, the algorithm performs favourably for the detection of the DME-0 class (AUC = 0.93) and the DME-2 class (AUC = 0.97), and these results are similar when compared to the second ophthalmologist using the test data for DME-0 (AUC = 0.89) and for DME-2 (AUC = 0.96). The DME-1 class is less well detected when compared to both the first ophthalmologist (AUC = 0.83) and the second ophthalmologist (AUC = 0.75). However, it can be noted that the DME-1 class also has the lowest number of samples, which could affect the performance. It can also be noted that as per the DME detection phase, variability between the ophthalmologist gradings may affect the algorithm predictions, and it can be noted that the algorithm performs more favourably when compared to the first ophthalmologist gradings. This situation is similar for the testing of DME detection and could be improved in future work by having an increased number of graders involved in the training stage and using an increased number of images, particularly for the DME-1 class. Furthermore, discussing difference in assigned grades would establish a more robust gold standard.

Overall, the algorithms in this study have demonstrated the ability to classify DME, and there are notably low rates of error for the DME-2 class, which would be ideal for use in screening applications. In the previous similar study by Tang et al., they achieved an AUC of 0.958 with their primary dataset, and AUC results between 0.936 and 0.956 for their external datasets using images from the Heidelberg Spectralis machine for the detection of DME.[24] They achieved an AUC of 0.951 for the distinguishing of CI-DME versus non-CI-DME with their primary dataset, and AUC scores between 0.899 and 0.934 for their external datasets. Thus, the overall AUC scores achieved in this study are similar and even though the accuracy scores are higher for the study completed by Tang et al., the datasets used have many images taken from the same patients, unlike this study which had no patient duplication and may have more heterogenous data. Despite the images also being classified in a slightly different manner in the previous study, this work supports the previous finding that an automated classification of DME level from OCT scans is feasible.

This has practical implications and can be used to direct treatment decisions and prevent blindness due to DME, which is relevant as the number of patients with DM continues to grow worldwide.[1] The ICO has guidelines for the management of DME based on the location on the retina,[2] and the Scottish diabetic retinal screening service already use DME levels in retinal photographs to direct treatment or referral.[33] While macular

thickening may not always be associated with a change in the patient's vision,[25] reduced visual acuity is associated with a higher prevalence of DME,[4] and thus combining visual acuity with the level DME as has been previously suggested may represent a suitable implementation of automated DME severity classification in screening programmes.[2] A reduction in vision in conjunction with the presence of DME should indicate the most urgent referral, and using such software in a screening service could expedite treatment.

There are several limitations to consider for this study. There was a limited variety in image grading, and the gold standard for the training and validation images was set by a single ophthalmologist. Future work should use several gradings at this stage, with discussion on any disagreement. The amount of data was also limited in this study due to time constraints. There is a population bias as the images were only taken from a hospital ophthalmology clinic. While an automated approach could be used for the detection of the fovea,[34] there could be signs of DME elsewhere in the eye and only images at the foveal level were included. Furthermore, only a single model of OCT machine was used, and further testing should use other machines to ensure applicability. The exclusion criteria were applied to simplify the approach, but this does not reflect the full extent of retinal disease that would be encountered in a screening service and the prediction ability may be affected by using images of differing quality. It is also worth noting that the algorithms in this test show lower sensitivity compared to specificity for the test data for DME detection and classification and could lead to increased missed cases over un-needed referrals. This may be less ideal depending on the requirements of a given screening programme, but further work could address this potentially using more sophisticated DL techniques. As clinicians and patients alike expect that AI systems should be reliable,[35] it is also important to consider how algorithms can provide a rationale for their prediction. This can be achieved by highlighting the most influential area on a scan,[35] and could also be implemented in future work.

CONCLUSIONS

This study has shown that DME severity can be classified from OCT scan images using DL as per the novel scale proposed. This supports the findings of previous similar work and adds to the evidence for the automated classification of DME severity. Such a system has clinical significance to speed up referral times for those in need of urgent treatment without compromising patient safety, and to effectively use resources. Further work could likely improve the classification performance.

ADDITIONAL STATEMENTS

Ethics: Ethics approval for this study was obtained from the Clinical Research Ethics Committee at Galway University Hospitals with reference number C.A. 2831. Informed consent was not required from patient due to the retrospective nature of the study, and the use of fully anonymised images. This study was not registered with any other body.

Funding: Funding was provided from a Health Research Board Summer Scholarship to the first author with reference number SS-2022-041.

Contributors: DOK, FH and CB proposed the study. DOK, FH, CB and AS contributed to the study design and methods, including the analysis. CB and RH collected and organised the data. FH and DT graded the images. CB wrote the software. CB prepared the original draft manuscript. DOK, FH, DT, RH, AS reviewed the manuscript. All authors have read and agreed to the final version of the manuscript.

Competing interests: The authors declare no competing interests.

Data availability statement: Data are available on reasonable request. Project code with the prediction data is referenced in the supplemental material and are available online.

REFERENCES

1 Tan GS, Cheung N, Simó R, *et al.* Diabetic macular oedema. *Lancet Diabetes Endocrinol*. 2017;5:143–55. doi: 10.1016/S2213-8587(16)30052-3

2 Wong TY, Sun J, Kawasaki R, *et al.* Guidelines on Diabetic Eye Care: The International Council of Ophthalmology Recommendations for Screening, Follow-up, Referral, and Treatment Based on Resource Settings. *Ophthalmology*. 2018;125:1608–22. doi: 10.1016/j.ophtha.2018.04.007

3 Goh JKH, Cheung CY, Sim SS, *et al.* Retinal Imaging Techniques for Diabetic Retinopathy Screening. *J Diabetes Sci Technol*. 2016;10:282–94. doi: 10.1177/1932296816629491

4 Olson J, Sharp P, Goatman K, *et al.* Improving the economic value of photographic screening for optical coherence tomography-detectable macular oedema: a prospective, multicentre, UK study. *Health Technol Assess*. 2013;17:1–142. doi: 10.3310/hta17510

5 Wong RL, Tsang CW, Wong DS, *et al.* Are we making good use of our public resources? The false-positive rate of screening by fundus photography for diabetic macular oedema. *Hong Kong Med J*. 2017;23:356–64. doi: 10.12809/hkmj166078

6 Virgili G, Menchini F, Casazza G, *et al.* Optical coherence tomography (OCT) for detection of macular oedema in patients with diabetic retinopathy. *Cochrane Database Syst Rev*. 2015;2015:CD008081.

7 Grauslund J, Andersen N, Andresen J, *et al.* Evidence-based Danish guidelines for screening of diabetic retinopathy. *Acta Ophthalmol*. 2018;96:763–9. doi: 10.1111/aos.13936

8 Pandey R, Morgan MM, Murphy C, *et al.* Irish National Diabetic RetinaScreen Programme: report on five rounds of retinopathy screening and screen-positive referrals. (INDEAR study report no. 1). *Br J Ophthalmol*. 2022;106:409–14. doi: 10.1136/bjophthalmol-2020-317508

9 Li D, Ran AR, Cheung CY, *et al.* Deep learning in optical coherence tomography: Where are the gaps? *Clin Exp Ophthalmol*. Published Online First: 2023. doi: 10.1111/ceo.14258

10 Ting DSW, Pasquale LR, Peng L, *et al.* Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol*. 2019;103:167–75. doi: 10.1136/bjophthalmol-2018-313173

11 Kermany DS, Goldbaum M, Cai W, *et al.* Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*. 2018;172:1122-1131.e9. doi: 10.1016/j.cell.2018.02.010

12 Li H-Y, Wang D-X, Dong L, *et al.* Deep learning algorithms for detection of diabetic macular edema in OCT images: A systematic review and meta-analysis. *Eur J Ophthalmol*. 2023;33(1):278–90. doi: 10.1177/11206721221094786

13 A p S, Kar S, S G, *et al.* OctNET: A Lightweight CNN for Retinal Disease Classification from Optical Coherence Tomography Images. *Comput Methods Programs Biomed*. 2021;200:105877. doi: 10.1016/j.cmpb.2020.105877

14 Wu Q, Zhang B, Hu Y, *et al.* Detection of morphologic patterns of diabetic macular edema using a deep learning approach based on optical coherence tomography images. *Retina*. 2021;41:1110–7. doi: 10.1097/IAE.0000000000002992

15 Rasti R, Rabbani H, Mehridehnavi A, *et al.* Macular OCT Classification Using a Multi-Scale Convolutional Neural Network Ensemble. *IEEE Trans Med Imaging*. 2018;37:1024–34. doi: 10.1109/TMI.2017.2780115

16 Asif S, Amjad K, Qurrat-ul-Ain. Deep Residual Network for Diagnosis of Retinal Diseases Using Optical Coherence Tomography Images. *Interdiscip Sci*. 2022;14(4):906–16. doi: 10.1007/s12539-022-00533-z

17 Lu W, Tong Y, Yu Y, *et al.* Deep Learning-Based Automated Classification of Multi-Categorical Abnormalities From Optical Coherence Tomography Images. *Transl Vis Sci Technol*. 2018;7:41. doi: 10.1167/tvst.7.6.41

18 Wang X, Tang F, Chen H, *et al.* UD-MIL: Uncertainty-Driven Deep Multiple Instance Learning for OCT Image Classification. *IEEE J Biomed Health Inform*. 2020;24:3431–42. doi: 10.1109/JBHI.2020.2983730

19 Perdomo O, Rios H, Rodríguez FJ, *et al.* Classification of diabetes-related retinal diseases using a deep learning approach in optical coherence tomography. *Comput Methods Programs Biomed*. 2019;178:181–9. doi: 10.1016/j.cmpb.2019.06.016

20 Schlegl T, Waldstein SM, Bogunovic H, *et al.* Fully Automated Detection and Quantification of Macular Fluid in OCT Using Deep Learning. *Ophthalmology*. 2018;125:549–58. doi: 10.1016/j.ophtha.2017.10.031

21 Guo Y, Hormel TT, Xiong H, *et al.* Automated Segmentation of Retinal Fluid Volumes From Structural and Angiographic Optical Coherence Tomography Using Deep Learning. *Transl Vis Sci Technol*. 2020;9:54. doi: 10.1167/tvst.9.2.54

22 Lee CS, Tyring AJ, Deruyter NP, *et al.* Deep-learning based, automated segmentation of macular edema in optical coherence tomography. *Biomed Opt Express*. 2017;8:3440–8. doi: 10.1364/BOE.8.003440

23 De Fauw J, Ledsam JR, Romera-Paredes B, *et al.* Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*. 2018;24:1342–50. doi: 10.1038/s41591-018-0107-6

24 Tang F, Wang X, Ran A, *et al.* A Multitask Deep-Learning System to Classify Diabetic Macular Edema for Different Optical Coherence Tomography Devices: A Multicenter Analysis. *Diabetes Care*. 2021;44:2078–88. doi: 10.2337/dc20-3064

25 Browning DJ, Stewart MW, Lee C. Diabetic macular edema: Evidence-based management. *Indian J Ophthalmol*. 2018;66:1736–50. doi: 10.4103/ijo.IJO_1240_18

26 Panozzo G, Cicinelli MV, Augustin AJ, *et al.* An optical coherence tomography-based grading of diabetic maculopathy proposed by an international expert panel: The European School for Advanced Studies in Ophthalmology classification. *Eur J Ophthalmol*. 2020;30:8–18. doi: 10.1177/1120672119880394

27 Xu J-J, Zhou Y, Wei Q-J, *et al.* Three-dimensional diabetic macular edema thickness maps based on fluid segmentation and fovea detection using deep learning. *Int J Ophthalmol*. 2022;15:495–501. doi: 10.18240/ijo.2022.03.19

28 Kermany D, Zhang K, Goldbaum M. Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification. *Mendeley Data [Dataset]*. 2018;V2. doi: 10.17632/rscbjbr9sj.2

29 Cohen J. A Coefficient of Agreement for Nominal Scales. *Educ Psychol Meas*. 1960;20:37–46. doi: 10.1177/001316446002000104

30 Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. *Proceedings of the 3rd International Conference on Learning Representations*. San Diego, USA 2017.

31 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*. 1988;44:837–45. doi: 10.2307/2531595

32 Wilson EB. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*. Published Online First: 1 June 1927.

33 Zachariah S, Wykes W, Yorston D. Grading diabetic retinopathy (DR) using the Scottish grading protocol. *Community Eye Health*. 2015;28:72–3.

34 Schurer-Waldheim S, Seebock P, Bogunovic H, *et al.* Robust Fovea Detection in Retinal OCT Imaging using Deep Learning. *IEEE J Biomed Health Inform*. 2022;26:3927–37. doi: 10.1109/JBHI.2022.3166068

35 Young AT, Amara D, Bhattacharya A, *et al.* Patient and general public attitudes towards clinical artificial intelligence: a mixed methods systematic review. *Lancet Digit Health*. 2021;3:e599–611. doi: 10.1016/S2589-7500(21)00132-1