Artificial Intelligence in Diabetic Eye Disease Screening

Carol Y. Cheung, PhD, * Fangyao Tang, PhD, * Daniel Shu Wei Ting, MD, PhD; †
Gavin Siew Wei Tan, MD, † and Tien Yin Wong, MD, PhD †

**Abstract:** Systematic or national screening programs for diabetic retinopathy (DR) and diabetic macular edema (DME), using digital fundus photography and optical coherence tomography (OCT), are currently implemented at primary care level, aiming to provide timely referral for vision-threatening DR and DME to ophthalmologists for timely treatment and vision loss prevention. However, interpretation of retinal images requires specialized knowledge and expertise in diabetic eye disease. Furthermore, current DR screening programs are capital- and labor-intensive, which makes it difficult to rapidly scale up and expand diabetic eye screening to meet the needs of this growing global epidemic. Deep learning (DL), a new branch of machine learning technology under the broad term of artificial intelligence (AI), has made remarkable breakthrough in medical imaging in particular for pattern recognition and image classification. In ophthalmology, AI and DL technology has been developed from big image datasets in assessment of retinal photographs for detection and screening of DR as well as the segmentation and assessment of OCT images for diagnosis and screening of DME. This review aimed to summarize the current progress and the development of AI and DL technology for diabetic eye disease screening as well as current challenges in the actual implementation of DL in screening programs, and translating DL research into direct clinical applications of screening in a community setting.

**Key Words:** artificial intelligence, deep learning, diabetic retinopathy, optical coherence tomography, screening


Artificial intelligence (AI) using deep learning (DL) algorithms or systems has been identified as a major ‘disruptive innovation’ in medicine and healthcare. In ophthalmology, AI and DL technology has been developed in several areas, with the two most prominent being in assessment of retinal photographs for detection and screening of diabetic retinopathy (DR). 1–6

---

**WHY IT IS IMPORTANT TO SCREEN FOR DR?**

Screening for DR is based on a high level of evidence. By 2040, 642 million people worldwide is projected to have diabetes mellitus (DM). 22 This is a global epidemic with heavy health burden to individuals and societies across the world, affecting populations not only in highly developed countries, but also increasingly in developing countries. 23–24 Of note, DR is already the most specific microvascular complication of DM that leads to progressive visual impairment and even blindness. The prevalence of DR ranges from 30% to 45%, with 10% having vision-threatening DR (VTDR), defined as severe non-proliferative DR, proliferative DR, and diabetic macular edema (DME). 25–28 Diabetes is estimated to account for 11.6% of the annual healthcare budgets in most countries. 29 The cost of care for patients with DR is higher than those without DR, and increasing with severity. 30 With a projected increase in the incidence of DM, increased life expectancy and an aging population, it is anticipated that the impact of vision impairment associated with DR will inevitably increase in the coming years. Early identification of eyes at risk of VTDR and timely treatment has been shown to be effective in preventing the onset of visual impairment. 31 Recent results of randomized controlled trials have also shown that intraocular injections of drugs targeting vascular endothelial growth factor or inflammation in DME can lead to better visual outcomes. 31 Early detection through regular surveillance by clinical examination or grading of retinal photographs is, therefore, the key to prevent vision loss and even blindness.
CURRENT CHALLENGES IN DR SCREENING PROGRAMS

While the rationale to screen for DR among patients with DM is clear, there are significant challenges in designing and sustaining a DR screening program in many communities and countries.

Systematic or national screening programs for DR using 2-dimensional (2D) non-stereoscopic digital fundus photography are currently implemented at primary care level in many countries, aiming to provide timely referral for more severe levels of DR to ophthalmologists for timely treatment and vision loss prevention. Diagnosis of DME requires identification of thickening of the macula; therefore, screening for DME using non-stereoscopic fundus photographs is bound to a very high false-positive rate (eg, >86% in Hong Kong and >79% in the UK). Currently, spectral-domain optical coherence tomography (SD-OCT), which provides 3D volumetric data cube of the layered retinal structures, is being piloted to add in different DR screening programs in the UK, China and Singapore, aiming to reduce the false-positive rate of DME detection from fundus photography. When signs of referable DR or other referable eye diseases are identified from retinal photographs and SD-OCT, patients are referred to eye clinics at tertiary eye hospitals for clinical examination and management to prevent vision loss. However, interpretation of both retinal photographs and SD-OCT requires specialized knowledge and expertise in diabetic eye diseases and retinal imaging. Inter-individual differences in interpretation as well as potential of missing important retinal pathologies, especially with the engagement of non-ophthalmologist graders of varying skills and experience, are always concerned. Furthermore, current DR screening programs are capital- and labor-intensive, which makes it difficult to rapidly scale up and expand diabetic eye screening to meet the needs of this growing global epidemic. Given that the number of DM is expected to rise rapidly, manual grading of DR and DME from retinal images in DR screening programs will not be sustainable.

THE CASE FOR “AUTOMATED” MACHINE-BASED ASSESSMENT OF DR

In view of the challenges associated with DR screening, an automated program to assess DR from photographs has significant potential of increasing efficiency, reproducibility, and coverage of screening programs, and at the same time reducing cost and barriers to access, and improving patient outcomes by providing early detection and treatment.

Image interpretation using AI has been developed in the field of medicine. Over the last 2 decades, automated analysis of retinal photographs for DR has been studied using traditional machine learning methods based on feature extraction or pattern recognition specified by experts manually for DR characterization and classification. However, the detection of DR is a complex image-interpretation task. Based on feature extraction or pattern recognition using traditional machine learning algorithms, these systems still cannot reach a good level of specificity. Furthermore, traditional machine learning algorithms typically plateau in performance after it reaches a threshold of training data.

Recently, DL has been applied in the field of ophthalmology to outstrip the ceiling performance of prior technology, and has revolutionized the detection of DR/DME from retinal photographs and SD-OCT to achieve excellent diagnostic performance.

DL AS THE BREAKTHROUGH IN AI TECHNOLOGY

Deep learning is a new branch of machine learning technology under the broad term of AI. The concept is that DL permits software algorithms to be trained through a large mathematical function with millions of parameters on vast amounts of data called convolutional neural network (CNN). Such network is inspired by the ability of brains to learn complicated patterns in data by changing the strengths of synaptic connections between neurons. Deep learning uses deep networks with many intermediate layers of artificial “neurons” between the input and the output, and, like the visual cortex, these artificial neurons learn a hierarchy of progressively more complex feature detectors. In recent years, AI-DL with CNNs has been developed rapidly and has demonstrated strengths in intricate pattern recognition from big datasets in the medical field. For example, it is shown that the performance of automated classification for DR, retinal diseases, glaucoma as well as other medical conditions including malignant melanoma, tuberculosis, and lung cancer by DL systems are equal to or better than board-certified specialists. These studies underscore the ability of automated image interpretation using DL are promising.

For image interpretation, DL is superior to traditional machine learning. In traditional machine learning, features need to be extracted manually, before they are fed in the machine learning algorithm. Using DL, the features are automatically learnt by neural networks in the feature extraction stage and then fed into a classifier for classification. No lesion-based features need to be specified in DL, and thus steps of traditional feature calculation or lesion segmentation machine learning algorithms can be skipped which required considerable domain expertise and careful engineering. Furthermore, such DL approach may identify “different features” that are associated with the desired classification and unknown to humans previously. Table 1 summarizes recent DL systems for screening DR and DME. In the first approval from the FDA for detecting more-than-mild DR in adults who have DM using DL without assisted interpretation by a clinician. The case of IDx-DR highlights one of the earliest successes of a DL-based screening tool in completing the regulatory process.

© 2019 Asia-Pacific Academy of Ophthalmology

www.apjo.org | 159

Copyright © 2019 Asia-Pacific Academy of Ophthalmology. Unauthorized reproduction of this article is prohibited.
<table>
<thead>
<tr>
<th>Study</th>
<th>Year</th>
<th>Imaging</th>
<th>Output(s) of DR/DME</th>
<th>Deep Learning Techniques</th>
<th>Size of Dataset for Development</th>
<th>Size of Dataset for External Validation</th>
<th>Other Outputs</th>
<th>Sensitivity for Output of DR/DME</th>
<th>Specificity for Output of DR/DME</th>
<th>AUC for Output of DR/DME</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abrámoff et al&lt;sup&gt;6&lt;/sup&gt;</td>
<td>2016, 2018</td>
<td>Fundus photograph</td>
<td>mtmDR (defined as ETDRS level ≥35 and/or DME present)</td>
<td>AlexNet and VGG network architectures</td>
<td>Messidor-2: 1748 images</td>
<td>819 subjects recruited from 10 primary care sites</td>
<td>Sufficient image quality: yes/no</td>
<td>87.2%</td>
<td>90.7%</td>
<td>Not reported</td>
</tr>
<tr>
<td>Gulshan et al&lt;sup&gt;7&lt;/sup&gt;</td>
<td>2016</td>
<td>Fundus photograph</td>
<td>Referable DR</td>
<td>Inception-V3 architecture</td>
<td>128,175 images</td>
<td>EyePACS-1: 9963 images</td>
<td>Messidor-2: 1748 images</td>
<td>All-cause referable</td>
<td>EyePACS-1: 97.5%</td>
<td>EyePACS-1: 93.4%</td>
</tr>
<tr>
<td>Gargeya and Long&lt;sup&gt;8&lt;/sup&gt;</td>
<td>2017</td>
<td>Fundus photograph</td>
<td>Any stage of DR, mild DR</td>
<td>Customized CNNs</td>
<td>Messidor-2: 1748 images</td>
<td>Nil</td>
<td>Messidor-2 (no DR vs any stage of DR): 93%</td>
<td>376</td>
<td>0.975</td>
<td>0.961</td>
</tr>
<tr>
<td>Ting et al&lt;sup&gt;1&lt;/sup&gt;</td>
<td>2017</td>
<td>Fundus photograph</td>
<td>Referable DR and VTDR</td>
<td>Adapted VGGNet architecture</td>
<td>DR: 76,370 images Possible glaucoma: 125,189 images AMD: 72,610 images</td>
<td>Primary: 71,896 images 10 additional datasets: 40,752 images</td>
<td>Referable possible glaucoma</td>
<td>Referable DR: 90.5%</td>
<td>Referable DR: 91.6%</td>
<td>Referable DR: 0.936</td>
</tr>
<tr>
<td>Li et al&lt;sup&gt;9&lt;/sup&gt;</td>
<td>2018</td>
<td>Fundus photograph</td>
<td>Vision-threatening referable DR</td>
<td>Inception V3 architecture</td>
<td>71,043 images</td>
<td>35,201 images</td>
<td>Nil</td>
<td>92.5%</td>
<td>98.5%</td>
<td>0.955</td>
</tr>
<tr>
<td>Kanagasingam et al&lt;sup&gt;10&lt;/sup&gt;</td>
<td>2018</td>
<td>Fundus photograph</td>
<td>DR</td>
<td>Inception V3 architecture</td>
<td>30,000 images</td>
<td>193 subjects</td>
<td>Nil</td>
<td>100%</td>
<td>92%</td>
<td>Nil</td>
</tr>
<tr>
<td>Kermany et al&lt;sup&gt;11&lt;/sup&gt;</td>
<td>2018</td>
<td>OCT</td>
<td>DME</td>
<td>Inception V3 architecture and transfer learning</td>
<td>108,312 images</td>
<td>1000 images</td>
<td>CNV; drusen; urgent referral</td>
<td>Multiclass model: 97.8%</td>
<td>Multiclass model: 97.4%</td>
<td>Multiclass model: 0.999</td>
</tr>
<tr>
<td>De Fauw et al&lt;sup&gt;12&lt;/sup&gt;</td>
<td>2018</td>
<td>OCT</td>
<td>Diagnosis probability of macular edema</td>
<td>3D U-Net architecture for segmentation network Customized CNNs for classification network</td>
<td>877 segmented scans for the segmentation network; 14,884 scans with diagnoses and referral decisions for classification network</td>
<td>997 subjects</td>
<td>Referral suggestion, tissue volume of drusen and epiretinal membrane, and diagnosis probability of other retinal abnormalities</td>
<td>Not reported</td>
<td>Not reported</td>
<td>0.990</td>
</tr>
<tr>
<td>Li et al&lt;sup&gt;13&lt;/sup&gt;</td>
<td>2019</td>
<td>OCT</td>
<td>DME</td>
<td>VGG-16 network architecture</td>
<td>109,312 images</td>
<td>1000 images</td>
<td>CNV; drusen</td>
<td>98.8%</td>
<td>98.8%</td>
<td>0.999</td>
</tr>
</tbody>
</table>

AMD indicates age-related macular degeneration; AUC, area under the curve; CNN, convolutional neural network; CNV, choroidal neovascularization; DME, diabetic macular edema; DR, diabetic retinopathy; ETDRS, Early Treatment Diabetic Retinopathy Study; mtmDR, more-than-mild diabetic retinopathy; OCT, optical coherence tomography; VGG, Visual Geometry Group; VTDR, vision-threatening diabetic retinopathy.
2. A major development in DL for DR detection was the study by Gulshan et al1 from Google Healthcare which developed a DL system for detecting referable DR using 128,175 retinal photographs which were graded 3 to 7 times for DR, DME, and image gradability by a panel of 54 US-licensed ophthalmologists and ophthalmology senior residents. The DL algorithm achieved a high sensitivity (≥87%), specificity (≥90%) and area under the receiver operating characteristic curve (AUC) (≥0.99) in the external validation using 2 public databases (EyePACS-1: n = 9963 and Messidor-2: n = 1748); both graded by at least 7 US board-certified ophthalmologists.3 Other groups have reported similar results. Gargeya and Leng2 developed a DL algorithm for detecting DR (any DR and mild DR) using 75,137 retinal photographs obtained from the EyePACS public dataset. The DL algorithm has further been evaluated using 2 other public databases (Messidor-2: n = 1748 and E-Ophtal: n = 405) with a sensitivity of ≥74%, specificity of ≥80%, and AUC of ≥0.83.4 Li et al5 developed another DL algorithm and internally validated it using 71,043 retinal photographs acquired from a total of 36 hospital ophthalmology departments, optometry clinics, and screening settings in China. In the external validation, the DL algorithm was evaluated using 35,201 retinal photographs from population-based cohorts of Malay, Caucasian Australians, and Indigenous Australians with a sensitivity of 92.5%, a specificity of 98.5% and an AUC of 0.955.6 Kanagasingam et al7 developed a DL algorithm based on several training datasets including DiaRetDB1, EyePACS, and the Australian Tele-eye care DR database. They deployed the DL algorithm prospectively and evaluated the performance in 193 patients at a primary care clinic in Australia with a specificity of 92% and a positive predictive value of just 12% driven by the low incidence of DR in their cohort (2 of 193).8

3. A third major development addresses a major gap in previous studies in which the DL algorithms were not trained to detect other common sight-threatening eye conditions, such as glaucoma and AMD. To address this gap, Ting et al9 developed a DL system to screen for possible glaucoma and AMD, in addition to referable DR and VTDR. The DL system was composed of 8 CNNs, all using an adaptation of the VGGNet architecture: (i) an ensemble of 2 networks for the classification of DR severity; (ii) an ensemble of 2 networks for the identification of referable possible glaucoma; (iii) an ensemble of 2 networks for the identification of referable AMD; (iv) 1 network to assess image quality; and (v) 1 network to reject invalid non-retinal images. The DL system was validated using retinal photographs collected from the Singapore Integrated Diabetic Retinopathy Program (SIDRP) and from 10 additional multiethnic datasets from different countries with diverse community- and clinic-based populations with DM.1 In the primary validation dataset of 71,896 images from 17,974 patients, the AUC of the DL system was 0.936 for referable DR, 0.958 for VTDR, 0.941 for glaucoma suspect, and 0.931 for referable AMD. The DL system had AUCs between 0.889 and 0.983 for referable DR on 10 additional datasets (n = 40,752 images). In the secondary validation datasets, the DL system had AUCs between 0.889 and 0.983 for referable DR (n = 40,752 images). The DL system could also achieve high sensitivity (>90%), specificity (>73.3%) and AUC (>0.89) for identifying other referable eye conditions.

**DL IN OCT FOR DETECTING DME**

A major issue in DR screening is the detection of DME, which is not easily captured from 2-dimensional (2D) fundus photographs.3,5 Thus, recent studies have focused on demonstrating that DL algorithms can be trained using OCT images to detect DME and other retinal diseases. Kermany et al14 firstly applied DL and transfer learning techniques in the detection of DME as well as choroidal neovascularization, drusen, and normal from 2D OCT images. Two models (multiclass model and binary model) were firstly developed using a training dataset of 108,312 2D OCT images and a “limited model” was further trained using 1000 2D OCT images randomly selected from each class during training to compare transfer learning performance using limited data compared with results using a large dataset.14 In the validation, they evaluated the performance of the models of a dataset of 1000 2D OCT images and found that both models achieved high performance even in the limited model (sensitivity ≥96%, specificity ≥94%, and AUC ≥0.99).14 This study highlighted the application of transfer learning in conditions with small datasets.

Another major study was performed by Google’s Deepmind in collaboration with Moorfields Eye Hospital, where the group applied a combined 3D segmentation and classification networks for interpreting 3D OCT images in predicting tissue volumes and identifying and referring different retinal abnormalities.15 They developed the DL algorithm using 877 segmented scans for the segmentation network and 14,884 scans from 7621 patients with multiple clinical diagnoses and referral decisions for classification network. The classification network learned to map a segmentation map to 4 referral decisions (urgent, semi-urgent, routine, and observation only) and 10 different retinal pathologies (macular retinal edema, choroidal neovascularization, full-thickness macular hole, partial-thickness macular hole, epiretinal membrane, geographic atrophy, drusen, vitreomacular traction, central serous retinopathy, and normal). Their framework (predicting tissue volumes, identifying retinal abnormalities, and providing referral suggestions) closely matched the clinical decision-making process, separating judgments about the scan itself from the subsequent referral decision, allowing a clinician to inspect and visualize an interpretable segmentation, rather than simply being presented with a diagnosis and referral suggestion.

**VISUALIZING DL AND EXPLAINABILITY**

A key challenge for clinical adoption of DL algorithms is a major mindset shift in how clinicians entrust clinical care to machines. The issue of ‘black box’ problem has been identified as an impediment to the application of DL in healthcare.53 Previous reports have already performed visualization heat map analysis to represent intuitively the learning procedure of the CNNs for DR and DME.53 Recently, Keel et al54 conducted a study to analyze the CNN learning procedure for 2 validated DL systems including one for the detection of referable DR. They developed and applied a CNN-independent adaptive kernel visualization technique for visualizing the learning procedure of the networks. Using this method, discriminative image regions were highlighted when the
classification possibility output of being diagnosed was greater than 70%, and a heat map was then generated highlighting highly prognostic regions in the input image. They found that typical referable DR lesions (retinal hemorrhage, exudate, intraretinal microvascular abnormality, venous beading) could be identified as the most important prognostic regions in 96% of true-positive referable DR cases whereas nontraditional fundus regions (eg, optic disc and fundus areas adjacent to vessels) were identified in 85% of false-positive referable DR cases. This study may further promote the clinical adoption of using DL algorithm and enable clinicians to understand important exposure variables in real time.

HOW DOES DL FIT IN CURRENT DR SCREENING STRATEGY?

A critical question in the conceptualization of using DL technology for DR screening is “where does the DL system fit?” Deep learning systems could potentially be deployed in 2 different settings. Figure 1 illustrates the 2 proposed DL-based screening models for DR, compared with an existing model. First, the DL system could be incorporated in existing DR screening programs such as in the UK and Singapore to assist human assessors in centralized reading centers (a semi-automated model). The retinal images can be firstly analyzed by the DL system and the human assessors only review images flagged as “referable” or “ungradable” by the DL system (2nd screening) for further confirmation. This model not only decreases the massive workload to read those non-referable retinal images but also avoids those false-positive cases referring to ophthalmologists (Fig. 1A). Second, the DL system could be a fully automated model in that all retinal images can be analyzed by the DL system which will be useful in communities without any existing DR screening programs (Fig. 1B). By using this screening model, the sensitivity of the DL system may need to be set higher in order to minimize false-negative cases.

CURRENT CHALLENGES

The performance of DL algorithms is extremely promising as seen from current literature. However, there are still several challenges and complexities that limit the ability to move forward quickly.

First, as mentioned above, the “black box” approach is still unacceptable. Deep learning uses millions of image features most predictive for classification rather than explicitly detecting clinical features that physicians are familiar with (eg, microaneurysms, hard exudates). It is unclear exactly what the machine “thinks” and “sees”. Transparency of DL algorithm is required in order to let users understand the basis to justify a particular diagnosis, treatment recommendation, or outcome prediction that the DL algorithm offered, and also allow examination for any potential bias.

Second, AI most likely succeeds when it is used with high-quality “labeled” input data validated by multiple medical specialists for learning and classifying images in relation to outcomes. Also, DL algorithms are highly data hungry, often requiring millions of observations to reach acceptable performance levels. Moreover, training using single clinical dataset is always limited by potential biases (eg, same ethnicity, same device) which may affect both performance and generalizability.

Third, “real-world” experiments are essential in the

FIGURE 1. Two artificial intelligence–based screening models for diabetic eye diseases have been proposed. A, A semi-automated model that retinal images will be firstly analyzed by the deep learning systems. The images will be read by doctors or trained graders (2nd screening) if it is flagged by the deep learning systems. B, A fully automated model that retinal images will be analyzed by the deep learning systems. C, Existing model by human assessors.
validation of DL algorithm. For example, prospective study design in appropriate clinical settings and testing DL algorithm using independent validation datasets under different settings and conditions that played no role in model development are critical to evaluate the performance of DL algorithms. Several recent studies have already reported the performance of DL algorithm using a prospective study design.4,5,6

Fourth, a continuous data supply is necessary for ongoing improvement and refinement of the DL algorithms and data may need to be shared across multiple centers for widespread implementation. However, the current healthcare environment holds little incentive for data sharing.7

Lastly, before DL algorithms or systems can be deployed in the clinical workflow of diabetic eye disease screening programs, there are still more questions need to be addressed. For example, “If a patient suffers an adverse event due to an AI-based technology, who is responsible?”5 In addition, more rigorous scientific evidence on safety, validity, reproducibility, reliability, usability, and patient and clinician acceptability is also critical, before the clinical deployment.

CONCLUSIONS

In summary, AI and DL are entering the mainstream of clinical medicine. This technology can augment human intelligence to improve decision making and operational processes. Screening of DR is almost an ideal task for AI in health care. Looking forward, AI will become ubiquitous and indispensable for the screening, with expectation to improve the efficiency and accessibility of screening programs, and therefore will be able to prevent visual loss and blindness from this devastating disease.

REFERENCES


