Eye (2019) 33:97–109
https://doi.org/10.1038/s41433-018-0269-y

The ROYAL COLLEGE of OPHTHALMOLOGISTS

REVIEW ARTICLE

Fundus photograph-based deep learning algorithms in detecting diabetic retinopathy

Rajiv Raman[1] · Sangeetha Srinivasan[2] · Sunny Virmani[3] · Sobha Sivaprasad[4] · Chetan Rao[1] · Ramachandran Rajalakshmi[5]

Received: 25 September 2018 / Accepted: 7 October 2018 / Published online: 6 November 2018

Abstract

Remarkable advances in biomedical research have led to the generation of large amounts of data. Using artificial intelligence, it has become possible to extract meaningful information from large volumes of data, in a shorter frame of time, with very less human interference. In effect, convolutional neural networks (a deep learning method) have been taught to recognize pathological lesions from images. Diabetes has high morbidity, with millions of people who need to be screened for diabetic retinopathy (DR). Deep neural networks offer a great advantage of screening for DR from retinal images, in improved identification of DR lesions and risk factors for diseases, with high accuracy and reliability. This review aims to compare the current evidences on various deep learning models for diagnosis of diabetic retinopathy (DR).

Introduction

Proliferative diabetic retinopathy and diabetic macular edema are the two retinal sight threatening complications of diabetes. Screening and timely treatment of these complications have been shown to reduce blindness [1] due to these complications in the countries where these services are well established such as the United Kingdom. In England and Wales, patients with diabetes diagnosed by their General Practitioners are registered in a diabetes register and diabetic retinopathy screening services invite each patient for annual screening of the retina under mydriasis using standard cameras [2]. The images are graded systematically by trained trainers and images may be arbitrated if required. The re-call and referral pathways are also well-defined with minimal standards set for each severity grade of diabetic retinopathy.

Attempts to replicate these systematic diabetic retinopathy screening programs in low and middle income countries have not been successful. There are several challenges faced by these countries. First and foremost, the absolute numbers of people with diabetes and undiagnosed diabetes are significantly higher than in the United Kingdom. As an example, only about 4 million people with diabetes need to be screened annually for sight threatening complications in the United Kingdom. In contrast, there are over 70 million people with diabetes in India alone and an equal numbers of pre-diabetic or undiagnosed diabetes [3]. The primary care infrastructures of most low and medium income countries are in their infancy. Standard retinal cameras are too costly, electronic patient records are non-existent to develop a diabetes register and most importantly, there is significant shortage of trained personnel to capture the retinal images and grade them and to treat them. From the Asia-Pacific perspective, it is currently neither practical nor economical to have trained health care providers screen all 231 million (153 million and 78 million from Western Pacific and Southeast Asia, respectively) [4].

Therefore, there is an unmet and urgent need to develop bespoke clinical and cost-effective screening and treatment pathways that can cover at least the majority of the population with diabetes.

✉ Rajiv Raman
rajivpgraman@gmail.com

1 Shri Bhagwan Mahavir Vitreoretinal Services, Sankara Nethralaya, Chennai 600006, India

2 Vision Research Foundation, Chennai 600006, India

3 Verily Life Sciences LLC, South San Francisco, California, USA

4 NIHR Moorfields Biomedical Research Centre, London EC1V 2PD, UK

5 Dr. Mohan's Diabetes Specialities Centre and Madras Diabetes Research Foundation, Chennai 600086, India

Deep learning (DL) is a new Artificial Intelligence (AI) machine learning technique, and its use in the medical field has generated much interest over the last few years. DL mimics an infant's brain, which is like a sponge and learns through training. This technique can also be potentially used to detect diseases, as it can identify and classify data, image or a sound [5]. AI-chatbots with speech recognition capability have been explored to identify patterns in patient symptoms to form a potential diagnosis [6]. Researchers are using DL to train algorithms to recognize cancerous tissue comparable to trained physicians [7]. Likewise, the images from a fundus camera, microscope or radiography are being classified by DL and compared with the trained physician. Recently, DL has also been used to identify risk factors associated with cardiovascular diseases (e.g., blood pressure, smoking and body mass index) from retinal photographs [8].

Over the past 2 years, there are many evidences on the use of DL algorithms to identify diabetic retinopathy (DR) either a binary model or multi-classifier models. This review aims to compare the current evidences on various DL models for diagnosis of DR.

Overview of deep learning methods

DL algorithm is considered as a fourth industrial revolution. It is based on learning features from the data. It processes large amount of data and extracts meaningful patterns from them [9]. In deep neuronal learning, the convolutional neural networks (CNNs) learn to perform their tasks through repetition and self-correction. A CNN algorithm teaches itself by analysing a labeled training set of expert-graded images and provides a diagnostic output [10]. If the network's diagnosis is wrong, the algorithm adjusts its parameters suitably to decrease the error. The network repeats the process for every image, until the system's output agrees with that of the human expert graders. Once the algorithm optimizes itself, it is ready to work on unknown images. Deep neural networks can detect subtle changes, patterns or abnormalities that may be possibly at times be overlooked by human experts [11].

The DL architecture found to be most suitable for imaging data is that of the CNNs [12]. Such networks contain special type of layers that apply a mathematical filtering operation known as convolution, making the individual neuron process data only for its receptive subfield. As the input image is processed with successive convolutional layers of the network, the filters in the process get stacked together creating progressively more descriptive feature detectors. During training, these individual detectors are then adjusted to detect those specific image features that are required to solve a particular image recognition task. Trained with large annotated datasets, these CNNs allowed computers to start recognizing visual patterns [13].

The entire approach in DL in DR does not involve judgment of individual retinal lesions and the feature extraction process is entirely automatic. The core analysis engine of most of the AI software used for retinopathy detection contains DR analysis algorithms—those for image enhancement, interest region identification, descriptor computation and screening classification in conjunction with an ensemble of deep neural networks for classification tasks such as image quality detection, diabetic retinopathy severity classification and detection of diabetic macular edema (DMO). Referable DR (RDR) is the main parameter identified by most AI algorithms and is defined as the presence of (i) moderate non-proliferative DR (NPDR) or higher and/or (ii) CSMO. Sight threatening DR (STDR) is defined by the presence of severe NPDR, proliferative DR (PDR) and/ or DMO [14].

Deep learning algorithms in DR: current evidences

Different statistical measures are available to quantify the performance of the model. The performance is measured by: sensitivity, which measures how many positive samples were correctly identified and, specificity, which measures the proportion of correctly identified negative samples. The graphical plot of a receiver operating characteristic (ROC) curve is used to find a trade-off between them. Receiver operating characteristic curves make use of the fact that many methods of labeling the image (grading) generate probabilities of assigning an input data sample to each possible output label. By changing the probability threshold for a decision, the proportion between the positive and negative label outputs change and, in this way, either the sensitivity or specificity of the model increases. In order to measure the overall performance of the algorithm, independent of a specific threshold and application, the area under the ROC curve (AUC) is used. The value of AUC lies between 0.5, which corresponds to a random guess, and 1.0, which shows 100% of specificity and sensitivity. All the current evidences in DL use these measures to evaluate the performance.

One of the earliest studies [15, 16] on automatic detection of DR from color fundus photographs was by Abramoff et al. in 2008 [17]. It was a retrospective analysis done with non-mydriatic images from EyeCheck DR screening project. They were able to detect RDR with 84% sensitivity and 64% specificity. In 2013, Abramoff et al. [18] published the results of sensitivity and specificity of the

Iowa Detection Program (IDP) to detect RDR and found a high sensitivity of 96.8% and specificity of 59.4%. The area under the AUC was 0.937.

In 2015, Solanki et al. [19] used their EyeArt AI software with Messidor2 dataset. EyeArt screening sensitivity for RDR was 93.8%, the specificity was 72.2% and the AUC was 0.94. In their next study, they evaluated an automated estimation of microaneurysm (MA) turnover, a potential biomarker for risk of progression of DR the tool identified new and disappeared microaneurysms with 100% sensitivity.

Since 2012, a large number of commercially available software were developed by many companies for the automated detection of DR, known as automated retinal image analysis systems (ARIAS). Tufail et al. [20] conducted a study in 2013 and published in 2017 to evaluate these systems. Retinal images analysed by three automated retinal image analysis systems namely iGradingM (UK), Retmarker (Portugal) and EyeArt (USA) were compared to standard manual grading of DR by human graders/ophthalmologists. EyeArt and Retmarker have higher sensitivity for RDR than human graders.

Abramoff et al. [21] in 2016 showed in their study that the integration of CNN to an existing DR detection algorithm resulted in improved detection of RDR by minimising the number of false positives. Using the Messidor-2 data set, a sensitivity of 96.8%, and a specificity of 87% for RDR was obtained in their study. The specificity improved from 59.4 to 87% when compared with their previous study in 2013. This hybrid screening algorithm, known as the IDx-DR became the first commercially available AI device to get US Food Drug Administration (FDA) approval for DR screening in April 2018. The IDx-DR is able to detect RDR (more than mild [mtm] DR) with a sensitivity of 87.4% and a specificity of 89.5%.

In 2016, Gulshan et al. [22] reported the results of the Google DL DR validation study. The algorithm was trained using 128,175 macula-centered retinal images obtained from EyePACS in the United States and retinal images from eye hospitals in India. In the break-through major validation study of Google algorithm for DRdetection, Gulshan et al. reported a high sensitivity and specificity for RDR (sensitivity of 97.5% and specificity of 93.4% in the EYEPACS-1 and 96.1% sensitivity and 93.9% specificity for Messidor-2 set).

A study by Gargeya et al. [23] with another DL algorithm to detect all stages of DR, showed a sensitivity of 94% and a specificity of 98% for RDR, with an AUC of 0.97 with EyePACS. External validation was done on the MESSIDOR-2 and E-Ophtha datasets in this study. Their study focused on identification of mild NPDR and not just RDR.

The most recent major study that reported on validation of DL was by Ting et al. [24] in Singapore. Their study included multiple retinal images taken with conventional fundus cameras from multiethnic cohorts of people with diabetes and their algorithm showed a high sensitivity and specificity for identifying DR and other eye diseases such as age-related macular degeneration and glaucoma. The sensitivity and specificity for RDR was 90.5% and 91.6%, respectively and for STDR, the sensitivity was 100% and the specificity was 91.1% in their study.

Another breakthrough was the study from India that used EyeArt™, the AI software on Remidio Fundus on Phone (FOP) mydriatic smartphone-based retinal images and showed 95.8% sensitivity and 80.2% specificity for detection of any DR, and 99.1% sensitivity and 80.4% specificity in detecting STDR [25]. The sensitivity and specificity for RDR was 99.3% and 68.8%, respectively.

There are many other algorithms like automated retinal image analysis systems which are suitable for automated detection support for both retinal photography and optical coherence tomography (OCT). Pegasus, a retinal AI platform developed by Visulytix (London, UK), supports comprehensive image analysis with both fundus imaging and OCT of the macula. It screens for glaucoma and age-related macular degeneration simultaneously while providing the severity of DR with a referral commendation. It identifies specific pathological retinal lesions and hence also allows the user to open the AI black box.

Assessing ground truth in studies

Within the last few years, there has been an exponential growth in the number of studies published on automated methods, especially DL for DR classification. It is also becoming clear that different studies have implemented substantially different approaches in determining the reference standard ("ground truth") that was used to measure and report the performance of their algorithms. The methodology and the quality of the reference standard can have a significant impact on the performance of the algorithms and this makes it challenging to compare different algorithms based on the performances published in these studies.

The classification of an image between one of the five DR grades involves evaluation of subtle lesions on retinal images. Additionally, these retinal images may vary in quality resolution, color, among other characteristics based on the camera used to acquire and software used to visualize the image. This makes DR grading a challenging and subjective process resulting in significant intergrader variability, as has been demonstrated by several studies [26–29]. Studies generally have multiple graders that read the same

image and come up with a method to resolve the disagreement between the grades. One method includes assigning a senior grader as an arbitrator to resolve the disagreement among grades from other graders. A study [20] comparing different automated DR image assessment software, used a modified version of this technique to resolve any disagreement between human grades and automated software grades to come up with a reference standard. Another approach is to come up with a consensus grading outcome as the final reference standard grade. In this approach, the image is sequentially assigned to individual graders until a certain number (usually three) of consistent grading outcomes are achieved. A study [30] evaluating the feasibility of AI-based DR screening algorithm at endocrinology outpatient services used this approach. Another approach as employed by Gulshan et al. [22] is a simple majority decision where the reference standard grade is considered to be the one with which a plurality of graders agree from a group of 3 or more independent graders. Another method, usually known as "adjudication", is where a group of 3 or more graders first grade each image independently and then discuss all disagreements until they are resolved. Krause et al. [31] examined the variability in different methods of grading, definition of reference standards and their effects on building DL models for detection of diabetic eye disease. In their study, the adjudication process resulting in a consensus grade from multiple retina specialists provided a more rigorous and less noisy reference standard for DL algorithm development, which also led to superior algorithm performance. Their study shows an improvement in the algorithm AUC to 0.986 from 0.930 for predicting mild or worse DR when using adjudication as the reference standard compared to majority decision while increasing image resolution. The study also yielded other insights about the discussion precision above the level typically used in everyday clinical practice. Other reference standard definition methods such as arbitration, consensus and majority grade don't officially include a step where individual graders get to discuss the rationale for their initial grade. This may be an important step for graders to ensure there is a consistent understanding of the grading guidelines between them and can help resolve disagreement.

The study by Krause et al. [31] also recognized that, although precise, the adjudication process is time consuming, costly and may not be a very practical approach for grading thousands if not millions of images for training the DL algorithms. They demonstrated a more pragmatic approach of adjudicating a small subset (0.22% in their cases) of the image grades used for development to make a "tuning" set, which, combined with existing clinical grades for the training set, significantly improved model performance without the need for adjudicating the entire training set.

Development dataset size its implications

For DL algorithm development, the dataset is usually divided into a train set, a tune set and a validation (or test) set. The train set is used for training the model parameters and the tune is used for determining algorithm hyperparameters and other model optimization choices. Finally, the validation set is used for measuring the final model performance. A general and traditional rule of thumb that DL scientists have used for a dataset on the order of 10,000 images or less is to split up the dataset as 70/20/10 percent (train/tune/validation). With datasets sizes getting much bigger recently (millions of images in some cases), it might be an acceptable approach to have a much smaller split for the tune and validation set and assign the majority of the cohort to the train set. However, it is still critical to have reasonably sized tune and validation sets with appropriate representation of each of the output classes being considered since the performance of the algorithm will be measured against those sets. As discussed in the previous section, the quality and precision of reference standard for the tune and validation sets may be critical for achieving algorithm performance improvements.

The overall size of the dataset required for training DL algorithm is dependent on various factors, including but not limited to one's desired output, performance target, and variability of the input data and the labels. Starting with a general example, suppose the goal of the algorithm is to classify images into two categories: images with dogs and images with cats, then in such an example, the dataset required to achieve an acceptable performance might be much smaller than one needed to train an algorithm for a more difficult task, e.g., classifying images into different breeds of dogs. To extrapolate this to DR, an algorithm that only classifies images into RDR and non-referable for DR might require fewer training images than one classifying image into a five-point International Classification of Diabetic Retinopathy (ICDR) grades. In addition, if any of the desired output categories has a low prevalence in the development dataset, the algorithm might need more images for that category in order to perform well. For example, proliferative DR (PDR) has a low prevalence in the general diabetic population, and some of its manifestations could be very rare, so more PDR images might need to be supplied in the training dataset.

The desired performance for the algorithm also plays a critical role in determining the size of the dataset. It is important to set these performance targets based on how the

algorithm is expected to be used. For DR, this could be determined by the target user: eye care providers or primary care providers. It could also be influenced by where the algorithm is expected to be used (low resource settings or high resource settings), if it would be used for screening or diagnostics, and if the algorithm is expected to be used as a primary read device or an assistive tool.

Finally, the variability in development data could also lead to significant changes in the dataset size required for acceptable performance. In the case of fundus images, variability may be caused by usage of fundus cameras of different manufacturers, different models of the same manufacturer, and different image acquisition technique (field of view, field location, flash intensity, image resolution, camera operator etc.). Variability in the grades or labels for these images is also important to consider. If one is using existing clinical grades that come with the images, variability may come from the grading rubric that was used to grade those images. Even if different clinical sites grade use standard grading rubrics, there are often differences in how different they grade images because some may be grading for screening and others for diagnosis, treatment or management as their primary goal may be to have an effect on the overall DR grade. Having such insights about one's dataset helps in determining such variability. To compensate, dataset size may be increased, or potentially a smaller subset of images may be adjudicated for a high quality reference standard, which may help balance out some of the noise in the labels of the train set as discussed by Krause et al. [31].

Image quality and deep learning

For DR screening, a basic requirement is to be able to acquire an image of the patient's retina that shows the macula and optic disc with sufficient image resolution, contrast and fidelity. Retinal cameras have been used by ophthalmologists for decades to image, diagnose, and document retinal pathologies. The retinal camera technology has made several advancements over the last few decades with improvements including but not limited to better image resolution, non-mydriatic imaging, comfort usability for patient and operator, extent of the retina that can be visualized and reducing the cost of the device. There are even several smartphone camera attachments now available that demonstrate the feasibility of performing retinal imaging while compromising on field of view and image quality. Expansion of screening programs to remote and rural areas can benefit from such devices since they are usually inexpensive and portable. However, most of these devices require dilation to capture images of sufficient quality [32]. These devices are usually hand-held that can make patient alignment, fixation and focus more challenging compared with more traditional fundus cameras. The latter have internal and external fixation aids along with a chin rest and head rest to stabilize the patient for easier image acquisition. It is to be noted that some smartphone-based retinal cameras [25, 33] offer a much smaller field of view (~20°) compared to a 45° field offered by traditional desktop cameras. One of the smartphone based retinal camera devices has made an effort to overcome some of these limitations and has proposed a Fundus on Phone (FOP) device that fits right onto a slit lamp stand, providing the stability to the operator and patient. A study published comparing FOP to a traditional desktop based fundus camera showed that FOP had reasonable sensitivity and specificity in detecting DR of varying severity [34]. However, they also acknowledged in this single hospital study that the quality of the images of FOP system was not as good as that obtained from the conventional desktop camera. Several handheld non-mydriatic cameras (non-smartphone based) have also been available for retinal imaging. They also offer a relatively less expensive method of capturing fundus images compared to traditional non-mydriatic cameras and also offer similar field of view (40°–45°) without requiring dilation. These cameras do however still have the limitation of handheld operation resulting in instability for operator and patient, as mentioned earlier. A study [35] to validate one of these handheld cameras for DR concluded that such devices might be adequate for DR screening in primary care centers and mobile units; however, there is room for improvement in terms of image quality, ease of operation and patient comfort. Traditional desktop fundus cameras do address many of the limitations of the smartphone-based and handheld cameras, however they do have a larger footprint, making portability a challenge and can be cost prohibitive depending on where the care is being provided.

The Early Treatment Diabetic Retinopathy Study (ETDRS) group introduced stereoscopic color fundus photography in seven standard fields as the gold standard for the detection and classification of DR. It can be very time-consuming to capture a set of seven-field ETDRS images on traditional fundus cameras that require pupil dilation and a skilled operator. To address this, an ultrawide-field imaging device that can visualize the peripheral retina in a single capture without dilation, has been shown [36] to improve diagnosis and classification of DR.

For DL algorithms that can automatically classify fundus images for DR for screening purposes, the quality of images being input is key. Many of these DL algorithms are now trained on several thousands to even millions of images, and owing to a rich mix of images in this training data they are able to accept some noise, artefacts and misalignments. However, most algorithms also specify the minimum

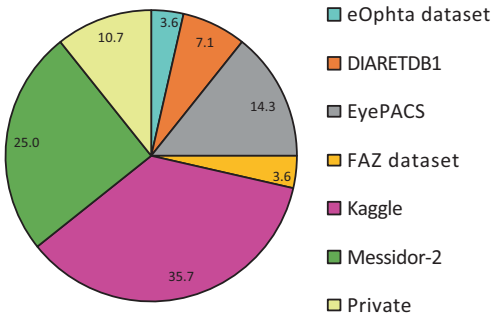

Fig. 1 Public datasets used % in the development and validation of deep learning algorithms ($n = 28$)

requirements in terms of image quality that will be acceptable for providing a diagnosis. One approach to convey image quality assessment is an image quality score (very commonly used in optical coherence tomography images, known as signal strength) and the second one may be to call an image of insufficient or sufficient quality for diagnosis by an algorithm similar to how it is done by IDx-DR [37]. It is important for groups that are setting up DR screening programs and plan to use automated DR classification methods to recognize that even though some cameras may offer inexpensive ways of capturing fundus images, if the rejection or ungradable rate for these images is too high, it could impact the success of the program. For all images that are deemed ungradable by the algorithm, it is prudent to recommend the patient to be seen by an eye-specialist since a referral/non-referral recommendation could not be made. If the rate of these ungradable images is high because of the quality of the retinal camera being used, the program may result in too many false positive referrals, leading to higher patient care costs, not to mention the frustration for the operator and patient if image retakes are needed.

Presence and absence of the disease and staging of the disease

The standardized diabetic retinopathy classification schemes enable a multidisciplinary approach where different medical specialists, including retina specialists, general ophthalmologists, optometrists, internists, endocrinologists, and family care physicians use the same language for giving an optimal care to the patient.

The AI systems have used binary information like presence or absence of DR [25, 38], and RDR or non-referable cases [17, 25] or multiple staging like the international clinical diabetic retinopathy disease severity scale for DR (ICDR) [14] and early treatment diabetic retinopathy study (ETDRS) [39] to classify the disease. For screening, both classifications have their own value. The five-point grading also provides information about urgency of referral rather than just RDR/no-referral. When the multiple staging is used, it is important to see that the sample size are calculated to power for individual stages, particularly the STDR.

Is there an effect of race/ethnicity or fundus pigmentation?

The density of the background pigmentation of the fundus oculi is different for different races. Some fundi are lightly pigmented often called "blonde" fundus and some are heavily pigmented called slate gray appearing fundus, both still within the range of normality. Although the lesions of DR are same across all races, the background color may make them simple or difficult to provide a confirmed diagnosis. Many of the AI algorithms have used Kaggle/Messidor/Eyepacs images for training them. Figure 1 shows public datasets used % in the development and validation of deep learning algorithms. These datasets are not representative of different races, thus the performance of the algorithm may differ across different races depending on pigmentation.

Ting et al. [24] reported the development and performance of their algorithm in multi-ethnic population and with different fundus camera. The DL system showed consistent results in darker fundus pigmentation in African American and Indian individuals to lighter fundus in white individuals.

Performances of deep learning algorithms for diabetic retinopathy

In DL, the performance of a model can be evaluated based on its prediction accuracy on separate test data samples which were not present in the training dataset. If the performance of a model is good on the training dataset but poor on the test dataset, the model has learned very specific patterns and is referred to as "overfitted" to the training data. A well-fitted model performs accurately on the training data and the test data. Table 1 compares various studies on DL for DR based on fundus photograph. Figure 2 shows the automated AI DR report of a normal retina (1 A) and for RDR/ STDR (IB)

The iDx-DR device combines an AI- and cloud-based algorithm with retinal fundus camera. Retinal images of sufficient quality are differentiated into negative (=non-referable = no or mild DR) or positive, indicative of a condition of more than mild DR resulting in the referral to an ophthalmologist (RDR) by the algorithm. The FDA approval was granted on the basis of a study of 900 patients with diabetes at ten primary care sites. The algorithm has correct identification of RDR in 87.4% of the individuals and a correct negative result in 89.5% [37]. However, the

Table 1 Comparison of performance of deep learning-based DR algorithms for retinal photographs

S. no.	Authors	Year	*N*	Public data base used	Reference standard	Grading method	Sensitivity	Specificity	AUC
1	Teng T et al. [52]	2002	NA	NS	NS	NS	NS	NS	NS
2	Arenas-Cavalli [53]	2015	450	NS	Ophthalmologist marked lesions/ features	NS	62.54%	91.89%	NS
3	Gupta S et al. [54]	2015	100	Messidor	Not specified	Not specified	87.00%	100%	NS
4	Bhaskaranand M et al. [55]	2016	40542	EyePACS	One Ophthalmologist	ICDR	90.00%	63.20%	0.879
5	Gulshan V et al. [22]	2016	139886	Messidor 2	Ophthalmologists majority consensus	ICDR	97.50%	93.40%	0.991
6	Pratt H et al. [56]	2016	80000	Kaggle	Kaggle grades	ICDR	30.00%- no DR vs any DR	95.00%- no DR vs any DR	NS
7	Pratumgul W et al. [57]	2016	600	NS	One Ophthalmologist	Normal, mild, moderate, severe	99.26%	97.77%	NS
8	Solanki K et al. [19]	2016	755	NS	NS	NS	90.50%	87.50%	0.965
9	Tufail A et al. [20]	2016	102856	NS	Manual Grade, Modified Arbitration	NHS DESP protocol	IGradingM-100%, Retmarker-85.00%, EyeNuk - 93.80%	iGradingM-0%, Retmarker-52.30%, EyeNuk-20.00%	NS
10	Walton B et al. [58]	2016	30030	PRIVATE	Optometrist or Ophthalmologist	ICDR	No DR x STDR: 66.40%	No DR x STDR: 72.80%	NS
11	Abbas Q et al. [59]	2017	750	DIARETDB1, FAZ, Messidor	NS	ICDR	92.18%	94.50%	0.924
12	Chandore V et al. [60]	2017	75000	Kaggle	Clinician rated	ICDR	88.88%	81.82%	NS
13	Dutta S et al. [61]	2017	1300	Kaggle	NS	NS	NS	NS	NS
14	Garcia G et al. [62]	2017	35126	EyePACS	EyePACS grades	ICDR	54.47%	93.65%	NS
15	Gargeya et al. [23]	2017	76885	EyePACS Messidor-2	Messidor-2 grades	NS	93.00%- No DR Vs Any DR	87.00%- No DR Vs Any DR	0.94
16	Lam C et al. [48]	2017	852	Kaggle,Tele Ophta	Two Ophthalmologists	Graded lesions	NS	NS	0.94: MA, 0.95: exudate
17	Lam C et al. [63]	2017	36555	Kaggle	NS	Normal, mild, moderate, severe	95.00%- Kaggle 90.00%-Messidor-2	96.00%- Kaggle 71.00%-Messidor	NS
18	Ling Dai et al. [64]	2017	646	DIARETDB1	Two Ophthalmologists	Graded MA	87.80%	96.10%	0.934
19	Raju M et al. [65]	2017	88252	Kaggle	Keggle	ICDR and laterality	80.28%	92.29%	NS
20	Rakhlin A et al. [66]	2017	NS	Kaggle and Messidor2	Messidor-2 and Kaggle grades	ICDR	99.00% -Messidor2 92.00% -Kaggle	71.00% -Messidor-2 72.00%-Kaggle	0.967-Messidor2 0.923-Kaggle
21	Takahashi H et al. [40]	2017	9939	NS	3 retinal specialists	Modified Davis grading		NS	NS

Table 1 (continued)

S. no.	Authors	Year	*N*	Public data base used	Reference standard	Grading method	Sensitivity	Specificity	AUC
22	Ting D et al. [24]	2017	297936	NS	2 senior nonmedical grades with extra 1 retina specialist for discordant cases	NS	Modified Davis grading-81% real prognosis- 96.00% rDR: 90.50% vtDR: 100%	rDR: 91.60% vtDR: 91.10%	rDR: 0.936 vtDR: 0.958
23	Torre J et al. [67]	2017	88650	EyePACS	NS	NS	NS	NS	NS
24	Xu X et al. [68]	2017	1000	Kaggle	One Ophthalmologist	HA, lesions; MA, blood vessels	NS	NS	NS
25	Bhattacharya S et al. [69]	2018	200	NS	NS	Normal X NPDR	89.09%	92.22%	NS
26	Desbiens J et al. [70]	2018	7000	Messidor-2	Ophthalmologist	ICDR	92.90%	98.90%	NS
27	Keel S et al. [30]	2018	58886	NS	Consensus grading	NHS DESP protocol	92.30%	93.70%	NS
28	Kermany D et al. [71]	2018	2000	NS	6 experts		96.80%	99.60%	NA
29	Krause J et al. [31]	2018	1605695	Messidor 2	Retina specialist face to face adjudication	ICDR/EyePACS	97.10%	92.30%	0.986
30	Kwasigroch A et al. [72]	2018	37000	NS	NS	ICDR	89.50%	50.50%	NS
31	Rajalakshmi R et al. [25]	2018	2048	NS	2 ophthalmologists with 3rd ophthalmologist as adjudicator	ICDR	No DR x DR: 95.80% No DR x STDR: 99.10%	No DR x DR:80.2% No DR x STDR: 80.4%	NS
32	Suriyal S et al. [73]	2018	16798	Kaggle	NS	DR and No DR	74.50%	63.00%	NS
33	Venugopal G et al. [74]	2018	NS	NS	NS	NS	86.30%	NS	NS
34	Voets M et al. [75]	2018	98679	Kaggle and Messidor2	One non-clinician	Moderate + NO DME	75.40%- Kaggle 57.60%-Messidor-2	55.40%-Kaggle 54.60%-Messidor-2	0.74-Kaggle 0.59-Messidor

NS not specified; *NHS DESP*, national health service diabetic eye screening program, *DR* diabetic retinopathy, *STDR* sight threatening diabetic retinopathy, *DIARETDB1* diabetic retinopathy database 1, *ICDR* international clinical diabetic retinopathy, *HA* Hemorrhages, *MA* microaneuryms, *DME* diabetic macular edema

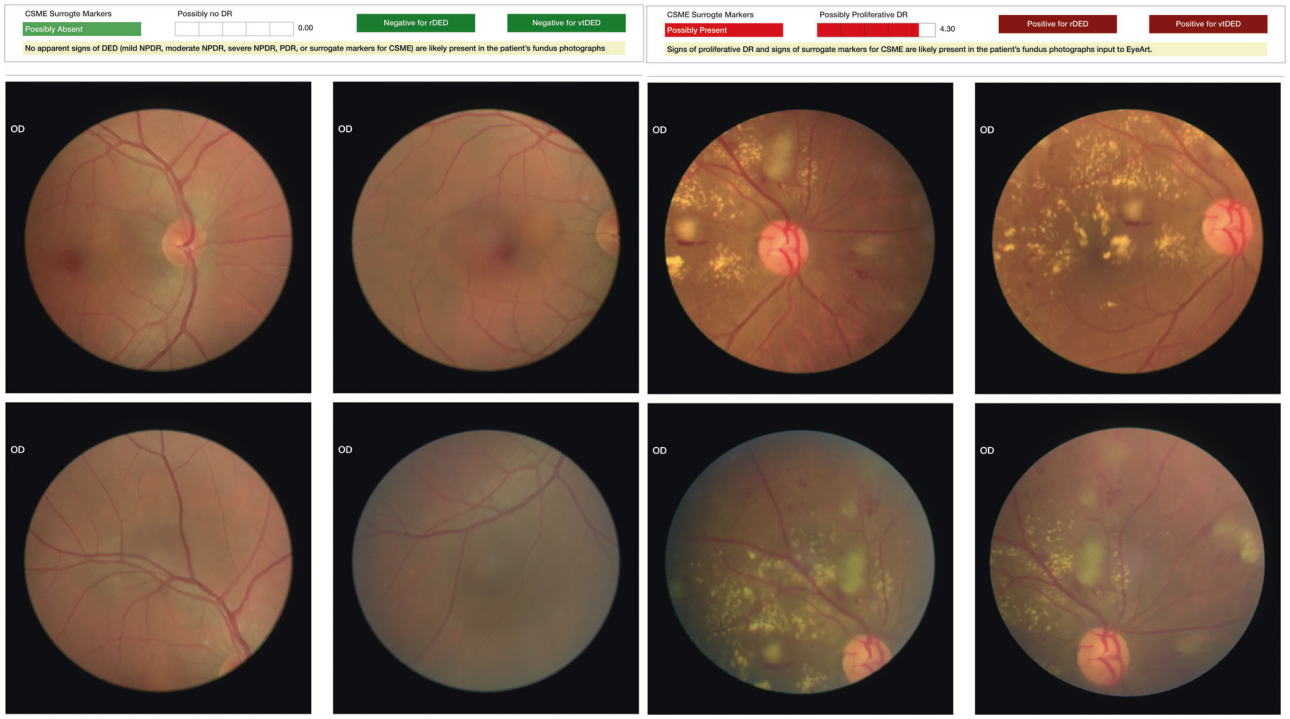

Fig. 2 Sample report of AI-based DR detection using EyeArt

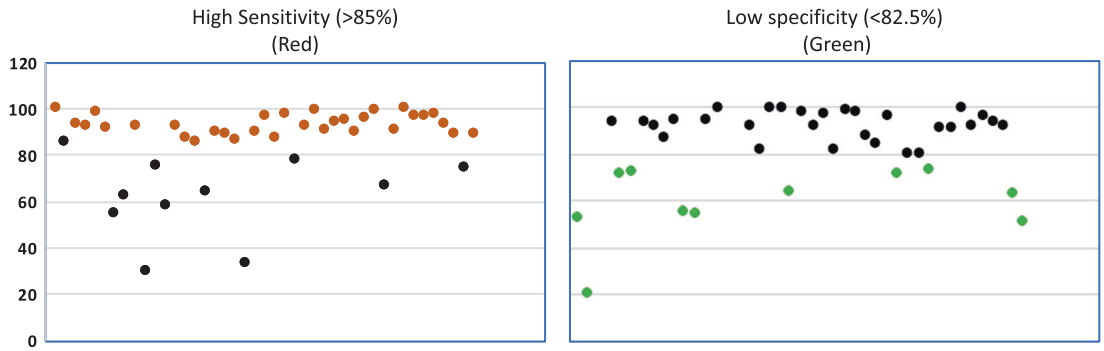

Fig. 3 Scatter plot showing sensitivity and specificity of various studies with cut-off point suggested for IDx algorithm

detection algorithm was trained on DR in untreated eye images, previous laser, pharmacological treatment or surgery were excluded. Moreover, those with manifest disease such as, severe NPDR or PDR were excluded. The FDA had set a mandatory level of accuracy as the primary endpoint for this trial with a sensitivity of more than 85% and a specificity of more than 82.5%. As it is intended to be a screening device, the FDA had chosen a higher sensitivity.

In other reported DR screening algorithms, the sensitivity varied from 87 to 97%, and the specificity from 59 to 98% (Table 1). So a majority of the available AI methods would be capable of being used for DR screening according to the requested FDA endpoints, and most of them seem to be performing better and faster than clinicians. Figure 3 shows the scatter-plot depicting the studies which meet the FDA cut-offs. However, a direct comparison of different algorithms is difficult because of significant differences in grading rubric, grader experience, image quality and most importantly reference standard.

Gulshan et al. [22] reported the diagnostic grading/staging on their two datasets (8788 images and 1745 images). For "moderate or worse DR only" the sensitivity was 90 and 87%, respectively, and the specificity was 98% in both datasets; for "severe or worse DR only" the sensitivity was 84 and 88% and specificity 99 and 98%; for "DMO only" the sensitivity was 91 and 90% and the specificity 98 and 99%, respectively. The performance was optimised by training on 60,000 images, and a further increase did not increase the performance of the algorithm. Thus with a large dataset, the algorithm could perform well on disease stages

as well. Likewise, Ting et al. [24] also confirmed using almost half a million images with a sensitivity and specificity for RDR versus sight-threatening diabetic only (excluding moderate DR) of 91%/92% versus 100%/91%.

Rather than taking clinical grades for DR, Takahashi et al. [40] focused on diagnostic grading/staging by using the ground truth of actual interventions (laser, injections, surgery, nothing) performed after an image was taken. They included 4709 images and categorical visual acuity changes (improved, stable, worsened) for training and tested the algorithm on 496 cases, reaching an accuracy of 96% in the prediction of interventions compared with three retina specialists who reached an accuracy of 92–93%. However, the false-negative rate – when the grade was "not requiring treatment" but treatment was actually needed – was 12%. The false positive rate – when the grade was "requiring treatment in the current visit" but treatment was actually not needed– was 65%. These high false positive rates can increase the patient load, thus limiting its utility.

Thus, AI-based DR screening algorithms have reached or may even outperform the level of accuracy of clinical experts.

Limitations and further advancements

Although the AI-based models have achieved high accuracy in diagnosis of DR, the current challenge is the clinical validation and real time deployment of these models in clinical practice. Most of the studies used training sets from homogenous population of a region or a publicly available dataset [21, 24, 41]. Diversifying the dataset, in terms of ethnicities, and camera to capture the images will address this challenge to some extents [42].

Data access and big data sharing are most important in DL as the neural networks are intrinsically "data hungry". The public availability of ImageNet in 2009 catalyzed AI and is still the base of retina-based image analyses [43]. Open access to scientific data including retinal images is an important area in the global research agenda. However, the medical data is fundamentally and legally different, thus posing a challenge for "open access".

Also, the questions on privacy protection are particularly sensitive in retinal imaging as anonymization is not completely achievable owing to the individual nature of the retinal vasculature which provides a fingerprint-like individual feature. Thus, it is not possible to completely anonymize any medical images (magnetic resonance imaging of the brain or ophthalmic images). Thus, it is now referred to "de-identification" or "de-personalization" of medical images

Also, the training and tuning datasets are often subjected to many variables such as field of view, magnification, image quality and artefacts. The most important aspect for accuracy and universality of an algorithm is the quality of ground truth. The more accurate and robust the ground truth is, better and more universal would be the algorithm. More evidence on methods of getting high quality ground-truth labels are required.

Almost all current systems of DL for DR are based on cross-sectional data. These algorithms cannot handle the time factor, such as the disease incidence and progression. Only few studies of AI for DR have demonstrated the power calculation, which is important for the independent dataset. Pre-determining the required operating threshold on training set should be calculated using the prevalence, type 1, type 2 errors, precision and confidence intervals at the least.

The other challenge, "elephant in the room" is the black-box phenomenon. In DL, it is challenging to understand how exactly a neural network reaches a particular decision, or to identify which exact features it utilizes. How can the results of AI-based algorithms be properly understood by clinicians and researchers? How can we ensure the reliability of algorithms, if we cannot understand how they operate? Potential solutions to this problem are multi-step algorithms that first detect certain clinically known features (using DL) and then predict or classify based on these features [44]. Researchers have been attempting to generate heat maps highlighting the regions of influence on the image which contributed to the algorithm conclusion [45, 46]. However, such maps are often challenging and difficult to interpret [47]. Sometimes, it may highlight an area, with no visible disease [35, 48]. A recent study [49] demonstrated how assistance (including heat maps) from a deep learning algorithm can improve accuracy of DR diagnosis and prevent underdiagnoses improving sensitivity with little to no loss of specificity.

The AI-DR screening systems have been developed and validated using 2-dimensional images and lack stereoscopic aspects, thus making identification of elevated lesion like retinal traction and vitreous traction challenging. Using the information from multimodal imaging in future AI-algorithms may potentially address this challenge. In addition, the medico-legal aspects and the regulatory approvals are different in various countries and settings, which also need to be addressed before its use in real clinical settings.

One of the most important challenges to the clinical adoption of AI-based technology is to understand how the patients perceive to entrust clinical care to machines. Keel et al. [30] studied the patient acceptability of AI based screening model in an endocrinology outpatient setting and reported that 96% of participants were satisfied or very satisfied with the automated screening model. If the AI systems gain widespread use, we will start identifying more people who need treatment. The health authorities have to plan for this anticipated increase in the volume of referrals in their health system. Finally, it is important to point out

that most AI-based applications in medicine are still in the translational stage and have not yet demonstrated their benefit in clinical trials.

Role of deep learning and DR screening and its future

Advances in mobile hardware for DL have enabled iPhone and Android smartphones to run AI diabetic retinopathy algorithm offline for medical imaging diagnostic support. Medios AI software for DR detection works offline on the iPhone and produces instant pdf reports highlighting the lesions (heatmaps). AI has been promising in classifying two-dimensional photographs of retinal diseases and relies on databases of annotated images. Recent novel DL architecture applied to three-dimensional optical coherence tomography (OCT) scans has shown makes excellent appropriate referral recommendation [50]. DL analysis of OCT scans for morphological variations in the scan, detection of intraretinal fluid or subretinal fluid, neovascularisation have started gaining great interest recently [51]. Research on AI assisted automated OCT analysis to assess OCT biomarkers to predict outcomes of treatment such as intravitreal injections for various retinal disorders is on-going.

The AI devices provides a screening decision without requiring an ophthalmologist to interpret the retinal images, hence can be used by physicians who may not normally be involved in eye care. The integration of AI into healthcare, would be expected to radically change clinical practice with more people getting screened for retinopathy. With the exponential increase in the prevalence of diabetes, AI can ease the pressure on the healthcare system, particularly in India and in other lower and middle income countries with large number of people with diabetes to be screened for retinopathy with limited resources. AI based software in ophthalmology provide an easier and more convenient option for the people to detect the disease at an early stage at the physician clinic and hence the patient satisfaction levels could be better. Automated DR screening methods can make the screening process more efficient, cost-effective, reproducible, and accessible. A multicentre study in India (SMART India) will evaluate the clinical and cost-effectiveness of automated DR screening in India. The study is funded by the Global Challenge Research Fund from the United Kingdom Research and Innovations (UKRI) with the aim of translating the results globally.

Conclusion

Automated analysis with integration of AI in DR detection can be safely implemented and appears to be a promising solution for screening large number of people with diabetes globally. DL has shown impressive results of high sensitivity and specificity in automated image analysis of fundus photographs. Artificial intelligence would act as an auxiliary part and a useful assistant in DR screening and provide diagnostic support and cannot replace the role of ophthalmologists in clinical diagnosis and management. Future development of AI technology will generate medical advances, and the physicians and ophthalmologists would need to learn how to utilize the technical advances with care. Robust diagnostic AI-based technology to automate DR screening would help increase referral of people with STDR and other retinal diseases to the ophthalmologist / retina specialist for further management and hence would aid in reducing visual impairment.

The ultimate approach and goal of the future to save healthcare resources and possibly be to incorporate AI into a retinal imaging device fundus cameras that can be used in various locations such as pharmacy, optical shop, etc to screen all people with diabetes.

Compliance with ethical standards

Conflict of interest Sunny Virmani is an employee of Verily Life Sciences LLC. The authors declare that they have no conflict of interest.

References

1. Early photocoagulation for diabetic retinopathy. ETDRS report number 9. Early treatment diabetic retinopathy study research group. Ophthalmology. 1991;98:766–85.
2. Scanlon PH. The English National Screening Programme for diabetic retinopathy 2003–16. Acta Diabetol. 2017;54:515–25.
3. IDF Diabetes Atlas. 2017. http://www.diabetesatlas.org/. Accessed 18 Sept. 2018.
4. International Diabetes Federation. IDF Diabetes Atlas. 8th Edn Bruss. Belg. http://www.diabetesatlas.org/. Accessed 18 Sept. 2018.
5. Boddapati V, Petef A, Rasmusson J, Lundberg L. Classifying environmental sounds using image recognition networks. Procedia Comput Sci. 2017;112:2048–56.
6. Voice-based chatbots—a revolution in customer relations. Capgemini Worldw. 2017. https://www.capgemini.com/resources/voice-based-chatbots-a-revolution-in-customer-relations/. Accessed 18 Sept. 2018.
7. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J. 2015;13:8–17.
8. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng. 2018;2:158–64.
9. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds). Advances in Neural Information Processing Systems 25. Curran Associates, Inc., Redhook, 2012, pp 1097–105.

10. Browne M, Ghidary SS. Convolutional neural networks for image processing: an application in robot vision. In: Gedeon T (Tom) D, Fung LCC (eds). AI 2003: Advances in Artificial Intelligence. Springer Berlin Heidelberg, 2003, pp 641–52.
11. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw. 2015;61:85–117.
12. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.
13. Nebauer C. Evaluation of convolutional neural networks for visual recognition. IEEE Trans Neural Netw. 1998;9:685–96.
14. Wilkinson CP, Ferris FL, Klein RE, Lee PP, Agardh CD, Davis M, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology. 2003;110:1677–82.
15. Ege BM, Hejlesen OK, Larsen OV, Møller K, Jennings B, Kerr D, et al. Screening for diabetic retinopathy using computer based image analysis and statistical classification. Comput Methods Prog Biomed. 2000;62:165–75.
16. Goldbaum M, Moezzi S, Taylor A, Chatterjee S, Boyd J, Hunter E, et al. Automated diagnosis and image understanding with object extraction, object classification, and inferencing in retinal images. In: in retinal images," in 1996 IEEE International Conference on Image Processing. 1996, pp 695–8.
17. Abràmoff MD, Niemeijer M, Suttorp-Schulten MSA, Viergever MA, Russell SR, Ginneken Bvan. Evaluation of a system for automatic detection of diabetic retinopathy from color fundus photographs in a large population of patients with diabetes. Diabetes Care. 2008;31:193–8.
18. Abràmoff MD, Folk JC, Han DP, Walker JD, Williams DF, Russell SR, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. JAMA Ophthalmol. 2013;131:351–7.
19. Solanki K, Ramachandra C, Bhat S, Bhaskaranand M, Nittala MG, Sadda SR. Eyeart: automated, high-throughput, image analysis for diabetic retinopathy screening. Invest Ophthalmol Vis Sci. 2015;56:1429–1429.
20. Tufail A, Rudisill C, Egan C, Kapetanakis VV, Salas-Vega S, Owen CG, et al. Automated diabetic retinopathy image assessment software: diagnostic accuracy and cost-effectiveness compared with human graders. Ophthalmology. 2017;124:343–51.
21. Abràmoff MD, Lou Y, Erginay A, Clarida W, Amelon R, Folk JC, et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. Invest Ophthalmol Vis Sci. 2016;57:5200–6.
22. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316:2402–10.
23. Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. Ophthalmology. 2017;124:962–9.
24. Ting DSW, Cheung CY-L, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA. 2017;318:2211–23.
25. Rajalakshmi R, Subashini R, Anjana RM, Mohan V. Automated diabetic retinopathy detection in smartphone-based fundus photography using artificial intelligence. Eye. 2018;32:1138–44.
26. Scott IU, Bressler NM, Bressler SB, Browning DJ, Chan CK, Danis RP, et al. Agreement between clinician and reading center gradings of diabetic retinopathy severity level at baseline in a phase 2 study of intravitreal bevacizumab for diabetic macular edema. Retin Phila Pa. 2008;28:36–40.
27. Li HK, Hubbard LD, Danis RP, Esquivel A, Florez-Arango JF, Ferrier NJ, et al. Digital versus film fundus photography for research grading of diabetic retinopathy severity. Invest Ophthalmol Vis Sci. 2010;51:5846–52.
28. Gangaputra S, Lovato JF, Hubbard L, Davis MD, Esser BA, Ambrosius WT, et al. Comparison of standardized clinical classification with fundus photograph grading for the assessment of diabetic retinopathy and diabetic macular edema severity. Retina Phila Pa. 2013:33. https://doi.org/10.1097/IAE.0b013e318286c952.
29. Ruamviboonsuk P, Teerasuwanajak K, Tiensuwan M, Yuttitham K. Interobserver agreement in the interpretation of single-field digital fundus images for diabetic retinopathy screening. Ophthalmology. 2006;113:826–32.
30. Keel S, Lee PY, Scheetz J, Li Z, Kotowicz MA, MacIsaac RJ, et al. Feasibility and patient acceptability of a novel artificial intelligence-based screening model for diabetic retinopathy at endocrinology outpatient services: a pilot study. Sci Rep. 2018;8:4330.
31. Krause J, Gulshan V, Rahimy E, Karth P, Widner K, Corrado GS, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology. 2018;125:1264–72.
32. Garg SJ. Applicability of smartphone-based screening programs. JAMA Ophthalmol. 2016;134:158–9.
33. Russo A, Morescalchi F, Costagliola C, Delcassi L, Semeraro F. Comparison of smartphone ophthalmoscopy with slit-lamp biomicroscopy for grading diabetic retinopathy. Am J Ophthalmol. 2015;159:360–.e1.
34. Rajalakshmi R, Arulmalar S, Usha M, Prathiba V, Kareemuddin KS, Anjana RM, et al. Validation of smartphone based retinal photography for diabetic retinopathy screening. PLoS One 2015;10:e0138285.
35. Quellec G, Bazin L, Cazuguel G, Delafoy I, Cochener B, Lamard M. Suitability of a low-cost, handheld, nonmydriatic retinograph for diabetic retinopathy diagnosis. Transl Vis Sci Technol. 2016;5:16–16.
36. Ghasemi Falavarjani K, Wang K, Khadamy J, Sadda SR. Ultra-wide-field imaging in diabetic retinopathy; an overview. J Curr Ophthalmol. 2016;28:57–60.
37. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. Npj Digit Med. 2018;1:39.
38. Abràmoff MD, Garvin MK, Sonka M. Retinal imaging and image analysis. IEEE Rev Biomed Eng. 2010;3:169–208.
39. Grading diabetic retinopathy from stereoscopic color fundus photographs—an extension of the modified Airlie House classification. ETDRS report number 10. Early treatment diabetic retinopathy study research group. Ophthalmology. 1991;98:786–806.
40. Takahashi H, Tampo H, Arai Y, Inoue Y, Kawashima H. Applying artificial intelligence to disease staging: deep learning for improved staging of diabetic retinopathy. PLoS ONE. 2017;12:e0179790.
41. Lee CS, Tyring AJ, Deruyter NP, Wu Y, Rokem A, Lee AY. Deep-learning based, automated segmentation of macular edema in optical coherence tomography. Biomed Opt Express. 2017;8:3440–8.
42. Jelinek HF, Rocha A, Carvalho T, Goldenstein S, Wainer J. Machine learning and pattern classification in identification of indigenous retinal pathology. In: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2011, pp 5951–4.
43. Deng J, Dong W, Socher R, Li L, Li K, Fei-fei L. Imagenet: a large-scale hierarchical image database. In: In CVPR. 2009.
44. Choi JY, Yoo TK, Seo JG, Kwak J, Um TT, Rim TH. Multi-categorical deep learning neural network to classify retinal

images: a pilot study employing small database. PLoS ONE. 2017;12:e0187336.
45. Baig F, Mehrotra M, Vo H, Wang F, Saltz J, Kurc T. Sparkgis: Efficient comparison and evaluation of algorithm results in tissue image analysis studies. Biomed Data Manag Graph Online Querying VLDB 2015 Workshop Big-OQ DMAH Waikoloa HI USA August 31-Sept 4 2015 Revis Sel Pap Int Conf Very Large Data Bases 41st 2015 Wai 2016;9579:134–46.
46. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deepnetworks. *ArXiv170301365 Cs*2017. http://arxiv.org/abs/1703.01365. Accessed 25 Sept. 2018.
47. Ramanishka V, Das A, Zhang J, Saenko K. Top-down visual saliency guided by captions. A*rXiv161207360 Cs* 2016. http://arxiv.org/abs/1612.07360. Accessed 18 Sept. 2018.
48. Lam C, Yu C, Huang L, Rubin D. Retinal lesion detection with deep learning using image patches. Invest Ophthalmol Vis Sci. 2018;59:590–6.
49. Sayres R, Taly A, Rahimy E, Blumer K, Coz D, Hammel N, et al. Assisted reads for diabetic retinopathy using a deep learning algorithm and integrated gradient explanation. Invest Ophthalmol Vis Sci. 2018;59:1227.
50. Mandal A. Google's DeepMind AI could soon be diagnosing eye conditions. News-Medicalnet. https://www.news-medical.net/news/20180814/Googles-DeepMind-AI-could-soon-be-diagnosing-eye-conditions.aspx. Accessed 25 Sept. 2018.
51. Schlegl T, Waldstein SM, Bogunovic H, Endstraßer F, Sadeghipour A, Philip A-M, et al. Fully automated detection and quantification of macular fluid in oct using deep learning. Ophthalmology. 2018;125:549–58.
52. Teng T, Lefley M, Claremont D. Progress towards automated diabetic ocular screening: a review of image analysis and intelligent systems for diabetic retinopathy. Med Biol Eng Comput. 2002;40:2–13.
53. Arenas-Cavalli JT, Ríos SA, Pola M, Donoso R. A web-based platform for automated diabetic retinopathy screening. Procedia Comput Sci. 2015;60:557–63.
54. Gupta S, Kar AmI. Diagnosis of diabetic retinopathy using machine learning. J Res Dev. 2015;3:1–6.
55. Bhaskaranand M, Ramachandra C, Bhat S, Cuadros J, Nittala MG, Sadda S, et al. Automated diabetic retinopathy screening and monitoring using retinal fundus image analysis. J Diabetes Sci Technol. 2016;10:254–61.
56. Pratt H, Coenen F, Broadbent DM, Harding SP, Zheng Y. Convolutional neural networks for diabetic retinopathy. Procedia Comput Sci. 2016;90:200–5.
57. Pratumgul W, Sa-ngiamvibool W. The prototype of computer-assisted for screening and identifying severity of diabetic retinopathy automatically from color fundus images for mhealth system in thailand. Procedia Comput Sci. 2016;86:457–60.
58. Walton OB, Garoon RB, Weng CY, Gross J, Young AK, Camero KA, et al. Evaluation of automated teleretinal screening program for diabetic retinopathy. JAMA Ophthalmol. 2016;134:204–9.
59. Abbas Q, Fondon I, Sarmiento A, Jiménez S, Alemany P. Automatic recognition of severity level for diagnosis of diabetic retinopathy using deep visual features. Med Biol Eng Comput. 2017;55:1959–74.
60. Chandore V, Asati S. Automatic detection of diabetic retinopathy using deep convolutional neural network. Int J Adv Res Ideas Innov Technol. 2017;3:633–41.
61. Dutta S, Manideep BC, Basha SM, Caytiles RD, NCSN. Iyengar. Classification of diabetic retinopathy images by using deep learning models. Int J Grid Distrib Comput. 2018;11:89–106.
62. García G, Gallardo J, Mauricio A, López J, Del Carpio C. Detection of diabetic retinopathy based on a convolutional neural network using retinal fundus images. In: Lintas A, Rovetta S, Verschure PFMJ, Villa AEP (eds). Artificial Neural Networks and Machine Learning—ICANN 2017. Springer International Publishing, Switzerland AG, 2017, pp 635–42.
63. Lam C, Yi D, Guo M, Lindsey T. Automated detection of diabetic retinopathy using deep learning. AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci. 2018;2017:147–55.
64. Dai L, Fang R, Li H, Hou X, Sheng B, Wu Q, et al. Clinical report guided retinal microaneurysm detection with multi-sieving deep learning. IEEE Trans Med Imaging. 2018;37:1149–61.
65. Raju M, Pagidimarri V, Barreto R, Kadam A, Kasivajjala V, Aswath A. Validation of smartphone based retinal photography for diabetic retinopathy screening. Stud Health Technol Inform. 2017;245:559–63.
66. Rakhlin A. Diabetic retinopathy detection through integration of deep learning classification framework. *bioRxiv* 2018:225508.
67. Torre J. 2017—A deep learning interpretable classifier for diabetic retinopathy disease grading.pdf. Google Docs. https://drive.google.com/file/d/1_XBCyiPBIkJYuzn5GmjD8IU6n-OGl3NO/view?usp=drive_open&usp=embed_facebook. Accessed 18 Sept. 2018.
68. Xu J, Ishikawa H, Wollstein G, Bilonick RA, Folio LS, Nadler Z, et al. Three-dimensional spectral-domain optical coherence tomography data analysis for glaucoma detection. PLoS ONE. 2013;8:e55476.
69. Bhattacharya S, Sehgal J, Issac A, Dutta MK, Burget R, Kolarik M. Computer vision method for grading of health of a fundus image on basis of presence of red lesions. In: 2018 41st International Conference on Telecommunications and Signal Processing (TSP). 2018, pp 1–6.
70. Desbiens J, Gupta S, Stevenson J, Alderman A, Trivedi A, Buehler P. Deep annotated learning, harmonic descriptors and automated diabetic retinopathy detection. 2018. https://openreview.net/forum?id=BkuKMztoG. Accessed 18 Sept. 2018.
71. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2018;172:1122–.e9.
72. Kwasigroch A, Jarzembinski B, Grochowski M. Deep CNN based decision support system for detection and assessing the stage of diabetic retinopathy. 2018 Int Interdiscip PhD Workshop IIPhDW 2018;1:111–6.
73. Suriyal S, Druzgalski C, Gautam K. Mobile assisted diabetic retinopathy detection using deep neural network. 2018 Glob Med Eng Phys Exch Am Health Care Exch GMEPEPAHCE 2018:1–4.
74. Venugopal G, Viswanathan R, Joseph R. How AI enhances & accelerates diabetic retinopathy detection. 2018.
75. Voets M, Møllersen K, Bongo LA. Replication study: Development and validation of deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. arXiv preprint arXiv:1803.04337. 2018. http://arxiv.org/abs/1803.04337. Accessed 18 Sept. 2018.