Computer Methods and Programs in Biomedicine 191 (2020) 105320

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: A systematic review and meta-analysis

Md Mohaimenul Islam[a,b,f,1], Hsuan-Chia Yang[a,b,f,1], Tahmina Nasrin Poly[a,b,f], Wen-Shan Jian[e,*], Yu-Chuan (Jack) Li[a,b,c,d,f,*]

[a] Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan
[b] International Center for Health Information Technology (ICHIT), Taipei Medical University, Taipei, Taiwan
[c] Department of Dermatology, Wan Fang Hospital, Taipei, Taiwan
[d] TMU Research Center of Cancer Translational Medicine, Taipei Medical University, Taipei, Taiwan
[e] School of Health Care Administration, Taipei Medical University, Taipei, Taiwan
[f] Research Center of Big Data and Meta-Analysis, Wan Fang Hospital, Taipei Medical University, Taipei, Taiwan

ARTICLE INFO

Article history:
Received 9 July 2019
Revised 30 December 2019
Accepted 6 January 2020

Keywords:
Diabetic retinopathy
Deep learning
Fundus photograph
Diabetic
Retinopathy

ABSTRACT

Background: Diabetic retinopathy (DR) is one of the leading causes of blindness globally. Earlier detection and timely treatment of DR are desirable to reduce the incidence and progression of vision loss. Currently, deep learning (DL) approaches have offered better performance in detecting DR from retinal fundus images. We, therefore, performed a systematic review with a meta-analysis of relevant studies to quantify the performance of DL algorithms for detecting DR.

Methods: A systematic literature search on EMBASE, PubMed, Google Scholar, Scopus was performed between January 1, 2000, and March 31, 2019. The search strategy was based on the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) reporting guidelines, and DL-based study design was mandatory for articles inclusion. Two independent authors screened abstracts and titles against inclusion and exclusion criteria. Data were extracted by two authors independently using a standard form and the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool was used for the risk of bias and applicability assessment.

Results: Twenty-three studies were included in the systematic review; 20 studies met inclusion criteria for the meta-analysis. The pooled area under the receiving operating curve (AUROC) of DR was 0.97 (95%CI: 0.95–0.98), sensitivity was 0.83 (95%CI: 0.83–0.83), and specificity was 0.92 (95%CI: 0.92–0.92). The positive- and negative-likelihood ratio were 14.11 (95%CI: 9.91–20.07), and 0.10 (95%CI: 0.07–0.16), respectively. Moreover, the diagnostic odds ratio for DL models was 136.83 (95%CI: 79.03–236.93). All the studies provided a DR-grading scale, a human grader (e.g. trained caregivers, ophthalmologists) as a reference standard.

Conclusion: The findings of our study showed that DL algorithms had high sensitivity and specificity for detecting referable DR from retinal fundus photographs. Applying a DL-based automated tool of assessing DR from color fundus images could provide an alternative solution to reduce misdiagnosis and improve workflow. A DL-based automated tool offers substantial benefits to reduce screening costs, accessibility to healthcare and ameliorate earlier treatments.

1. Introduction

1.1. Rationale

Diabetes mellitus (DM) is a major public health concern in the United States and globally [1,2]. The world health organization (WHO) estimated that the global prevalence of DM was approximately 8.8% (95% confidence interval 7.2–11.3%) in 2017 and is ex-

* Corresponding authors at: Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan.
E-mail addresses: jj@tmu.edu.tw (W.-S. Jian), jack@tmu.edu.tw (Y.-C. (Jack) Li).
[1] Equal contribution.

https://doi.org/10.1016/j.cmpb.2020.105320
0169-2607/

pected to further rise to 9.9% (95% CIs 7.5–12.7) by the year 2045 [3,4]. Several potential factors are responsible for an increased prevalence of DM such as lack of physical activity, obesity, sedentary lifestyle and lack of awareness [5]. DM is, however, associated with microvascular complications such as diabetic nephropathy, neuropathy, and retinopathy [6]. DR is the most common complication of DM patients and often remains undetected until it progresses to an advanced vision-threatening stage [7]. Earlier detection and prevention of DR progression are required to reduce the incidence and progression of blindness [8]. A multidisciplinary approach including regular eye examinations and management of modifiable risk factors (e.g. glycemic control, hypertension, and hyperlipidemia, etc.) are effective to mitigate vision loss up to 98% of cases [9,10].

The American Diabetes Association (ADA) recommends screening for retinopathy for type 1 diabetic patients within three to five years after onset while type 2 diabetic need to screen one year after onset [11]. The subsequent examination should be carried out every six months to the one-year basis of severity of DR. Despite having standard guidelines for screening, approximately 30 to 50 percent of patients with DM are still not getting benefits from these recommendations [12]. Recently, multiple studies investigated the use of DL systems in automated detection of DR and achieved excellent diagnostic performance in terms of high sensitivity and specificity. A DL-based screening tool has substantial ability to improve health care by reducing the reliance on manual work, utilizing limited resources. An automated screening tool, therefore, could be used by any physicians who are not specialist in eye care or any place where healthcare personnel is not available.

1.2. Goals of this investigation

We herein report the results of a comprehensive systematic review of DL algorithms studies that have investigated the performance of DL algorithms for automated detection of DR in retinal fundus photographs. Our primary objective was to gauge precisely the performance of DL methods for the automated identification of DR from color fundus images. Clarification of the promising performance of DL algorithms may have relevant clinical implications for diagnosis, treatment, and prevention of DR.

2. Background

DR screening program was first introduced in the 1980s and early 1990s [13,14]. DR screening program successfully reduced the incidence of DR related vision loss at that time. In Sweden, early diagnosis of DR was initiated by mobile fundus photography teams to reduce the incidence of DR related vision loss by an average of 47% per year over 5 years follow up period in 1990 [4]. Currently, the International Council of Ophthalmology (ICO) provided some criteria's for DR screening that includes retinal examination with either (a) direct or indirect ophthalmoscopy or (b) mydriatic or non-mydriatic fundus imaging with $\geq 30^{\circ}$ mono- or stereo photography [15]. To measure the degree of DR severity, fundus image grading is essential and careful manual grading of these fundus images by an ophthalmologist can be labor-intensive, time-consuming and subjective. An automated algorithm may play an important role in analyzing DR fundus images; although previously used algorithms generally relied on traditional approaches that were involved with manually selecting engineering images features for some classifiers methods such as random forest and support vector machine [16,17]. In the early '90s, Gardner and colleagues first introduced an artificial neural network (ANN) approach that was capable of automatically grading DR with 88% sensitivity and 83% specificity relative to an ophthalmologist [18]. DL a branch of artificial intelligent learns task-specific DR image features with a variety of images without counting on manual feature selection. Therefore, a convolutional neural network (CNN) has recently become a state-of-the-art solution in a wide range of DR problems. The use of CNN for the automatic diagnosis of DR has provided better performance (higher sensitivity and specificity).

2.1. Gold standards

The primary care for DR includes timely ophthalmic examination and screening of high-quality retinal photographs of patients and vigilant follow-up of these patients for better and earlier treatment. Several common techniques such as direct and indirect ophthalmoscopy, stereoscopic color film fundus photography, mydriatic or nonmydriatic digital color or monochromatic photography technique have been used to detect and classify DR. The gold standard method for the detection and classification of DR is stereoscopic color fundus photographs in 7 standard fields (30°) which is recommended by the Early Treatment DR Study (ETDRS) group [19]. This method can also accurately identify diabetic macular edema (DME) and subtle retinal neovascularization [20]. This method is more accurate and reproducible, but it needs lots of skilled healthcare personnel (photographers, photograph readers, and good photography processing equipment).

2.2. Definition and grading of DR

DR is usually classified into two main stages- a) early-stage like non-proliferative DR (NPDR) which could be mild, moderate or severe b) advance stages including proliferative DR (PDR) and maculopathy or DME (Fig. 1). Patients tend to have vision loss due to macular edema and PDR. DR is usually asymptomatic; therefore, early identification and earlier treatment could prevent blindness [5]. The WHO recommended that patients with diabetes should undergo screening for retinopathy base on the severity of the disease. The Scottish Grading protocol has recommended 5 grades such as grades of no DR (R_0), mild NPDR(R1), pre-PDR (R2 &R3) that included moderate and severe NPDR, and PDR (R4) (Table 1) [21].

2.3. Artificial neural network

An ANN is inspired by the human brain and designed to recognize patterns. ANN receives inputs and integrates the inputs with weights and activated by several activation functions when a defined condition is satisfied (Fig. 2). It is comprised of inputs, hidden layers, and outputs. The input layer receives input values (e.g. images) and gives the output values such as a value or class (e.g. number or class of DR). The layers between the input layer and the output layer are called hidden layers.

An artificial neuron is simply described by the following equation:

$$y = f\left(\sum_{i=1}^{n} w_i x_i\right) = f(net)$$

Where, w_i is the connection weight of node i, x_i is the input of node i, and f is the activation function, which is usually a threshold function or a sigmoid function such as

$$f(net) = z + \frac{1}{1 + \exp(-x.net + y)}$$

Rosenblatt [22] invented the perceptron which used a gradient descent-based learning algorithm. The perceptron is centered on a

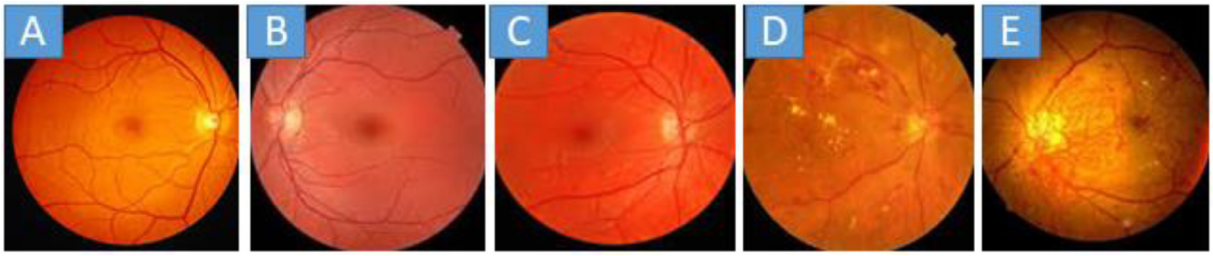


Fig. 1. Standard photographs from the patients with diabetic (A) No DR, (B) mild NPDR, (C) moderate NPDR, (D) severe NPDR, (E) PDR.

Table 1
The different grades of DR.

Grade	Retinopathy level	Clinical criteria	Outcome of screening
R0	Normal	No disease	Review in 12 months
R1	Mild NPDR	Micro-aneurysms, Flame exudate, >4 blot hemorrhages in one or both hemifields, and/ or cotton wool spots	Review range 6–12 months depending on the severity of signs, stability, systemic factors.
R2	Moderate NPDR	Venous beading, venous loop or reduplication. Intraretinal microvascular abnormality Blot hemorrhages	Review in approximately 6 months (PDR in up to 26%, high-risk PDR in up to 8% within a year)
R3	Severe NPDR	New vessels on disc (NVD) New vessels elsewhere (NVE) Pre retinal or vitreous haemorrhage Venous beading	Review in 4 months (PDR in up to 50%, high-risk PD is up to 15% within a year)
R4	PDR	New vessels on disc (NVD) New vessels elsewhere (NVE) Vitreous hemorrhage Retinal detachment	Refer ophthalmologist
M0	No maculopathy	No macular findings	12-month rescreening
M1	Maculopathy present	Hard exudate within 1–2 disc diameters of the fovea	6-month rescreening
M2	Referable maculopathy	Blot hemorrhage or hard exudates within 1 disc diameter of fovea	Refer ophthalmologist

Note: NPDR = Non-proliferative DR; PDR = Proliferative DR.

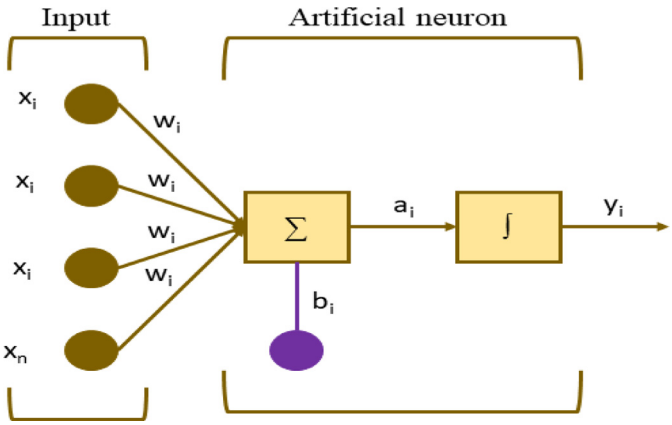


Fig. 2. Basic Framework of ANN.

single neuron and might be considered the primary basis of feed-forward ANNs. They use a simple step function and able to function as linear classifiers. The equation is given bellows:

$$f(net) = \begin{cases} 1\, if \sum_{i=0}^{n} w_i x_i > 0 \\ 0 \quad otherwise \end{cases}$$

Minsky and Papert [23] described several limitations regarding the perceptron, mentioning that they do not work on problems where the sample data are not linearly separable. The modern implementation of the backpropagation has a major advance on traditional gradient descent methods, in that it provided multi-layer feed-forward ANNs with a highly competitive supervised learning algorithm. The backpropagation algorithm is a supervised learning algorithm that helps network weights to try to minimize the Mean Squared Error (MSE) between the predicted outcome and the actual outcome of the network.

$$MSE = \frac{1}{P}\sum_{p=1}^{p}\sum_{j=1}^{k}\left(\left|O_{p,j} - d_{p,j}\right|\right)^2\right)$$

Where, $d_{p,j}$ is the predicted outcome, and $o_{p,j}$ is the actual outcome. In this process, weights between hidden and output nodes are modified by the following equation;

$$\Delta w_{k,j}^{2,1} = \eta\left(d_{p,k} - o_{p,k}\right)s'\left(net_{p,k}^{(2)}\right)x_{p,j}^{(1)}$$

Also, modify the weights between input and hidden nodes:

$$\Delta w_{j,i}^{1,0} = \eta\sum_{k}((d_{p,k} - o_{p,k})s'\left(net_{p,k}^{(2)}\right)w_{k,j}^{(2,1)})s'\left(net_{p,j}^{(1)}\right)x_{p,i}$$

2.4. Deep learning algorithms

DL is a more recently developed technique of machine learning, which mimics the human brain using multiple layers of ANN [24]. Although there are no explicit criteria on the threshold of depth to discriminate between shallow and DL, the latter is conventionally defined as having multiple hidden layers (Fig. 3).

Fig. 3 describes a $(L + 1)$ layer perceptron which is consists of N input units, O output units, and several so-called hidden units. A multiple perceptron consists of an input layer, an output layer, and L hidden layers. The i^{th} units within layer l calculate the output-

$$y_i^{(l)} = f\left(c_i^{(l)} \text{with} c_i^{(l)} = \sum_{k=1}^{m^{(l-1)}} w_{i,k}^{(l)} y_k^{(l-1)} + w_{i,0}^{(l)} \quad \text{(a)}\right.$$

Where $w_{i,k}^{(l)}$ denotes the weighted connection from the k^{th} units in layer $(l - 1)$ to the i^{th} units in layer l, and $w_{i,0}^{(l)}$ can be considered as an external input to the unit and is referred to as bias. Moreover, $m^{(l)}$ denotes the number of units in layer l, such that $N = m^0$ and $O = m^{(L+1)}$. Although, the bias can be regarded as a weight when introducing a dummy unit $y_0^{(l)} := 1$ in each layer.

$$c_i^{(l)} = \sum_{k=o}^{m^{(l-1)}} w_{i,k}^{(l)} y_k^{(l-1)} \quad \text{(b)}$$

Where c^l, w^l, and $y^{(l-1)}$ denote the corresponding vector and matrix represented of the actual inputs c_i^l, the weight $w_{i,k}^l$, and the outputs $y_k^{(l-1)}$, respectively.

The overall multilayer perceptron presents a function:

$$y(., w) : \mathbb{R}^N \to \mathbb{R}^O, x \to y(x, w)$$

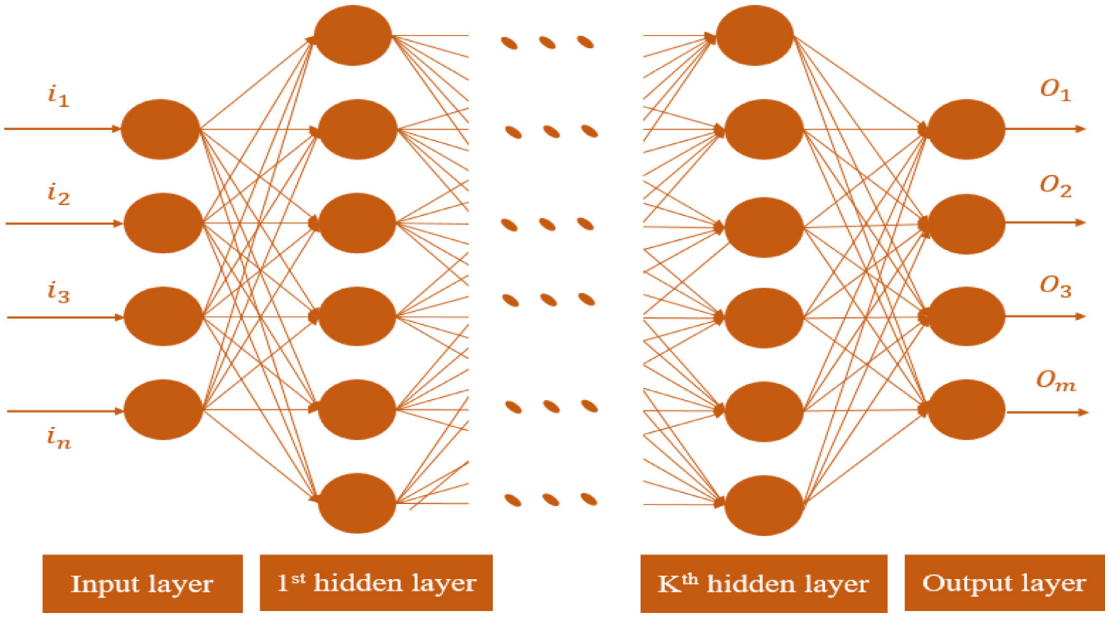


Fig. 3. Basic structure of DL.

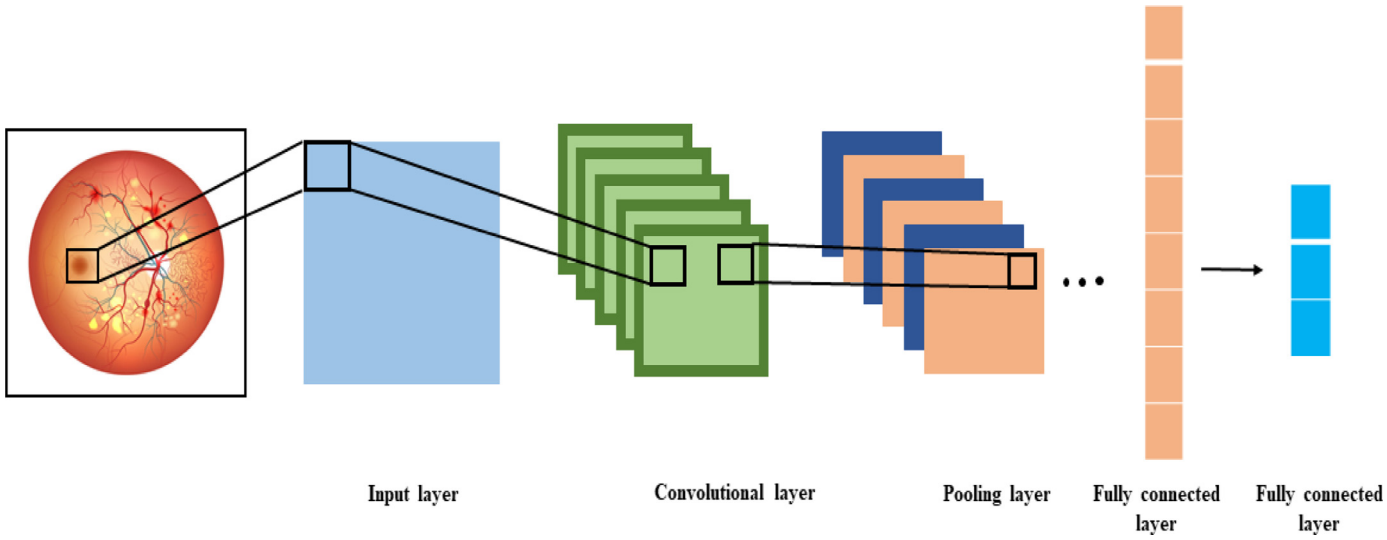


Fig. 4. Basic Framework of CNN.

2.5. Convolutional neural network

CNN is a DL technique which is widely used in interpreting medical images. CNN algorithms can identify faces, individuals, DR, pathology, tumor, and many other aspects of visual data. It takes and processes images as tensors and tensors are matrices of numbers with additional dimensions [25]. It consists of an input layer, convolutional layer, pooling layer, and fully connected layer (Fig. 4).

2.5.1. Convolutional layer

Suppose, layer l is a convolutional layer, and the input of layer l is comprised of $m_1^{(l-1)}$ feature maps from the previous layer, each of size $m_2^{(l-1)} \times m_3^{(l-1)}$. In the case, where $l = 1$, the input is a single image I consisting of one or more channels. CNN directly takes inputs from raw images. The output of layer l consists of $m_1^{(l)}$ feature maps of size $m_2^{(l)} \times m_3^{(l)}$. The i^{th} feature map in layer l denoted $y_i^{(l)}$ is calculated as

$$y_i^{(l)} = \sum_{j=1}^{m^{(l-1)}} k_{i,j}^{(l)} * y_j^{(l-1)} + b_i^{(l)} \quad (1)$$

Where $b_i^{(l)}$ is a bias matrix and $K_{i,j}^{(l)}$ is the filter of size $2h_1^{(l)} + 2h_2^{(l)} + 1$ connecting the j^{th} feature map in layer $(l-1)$ with the j^{th} feature map in layer l. As mentioned above, $m_2^{(l)}$ and $m_3^{(l)}$ are influenced by border effects. When applying the discrete convolutional only in the so-called valid region of the input feature maps, that is only for pixels where the sum of Eq. (1) is defined properly, the output feature maps have size

$$m_2^{(l)} = m_2^{(l-1)} - 2h_1^{(l)} \text{ and } m_3^{(l)} = m_3^{(l-1)} - 2h_2^{(l)} \quad (2)$$

Often the filters used for computing a fixed feature map $y_i^{(l)}$ are the same, that is $k_{i,j}^{(l)} = k_{i,k}^{(l)}$ for $j \neq k$. In addition, the sum in Eq. (1) may also run over a subset of the input feature maps. To relate the convolutional layer and its operation as defined by

Eq. (1) to the multilayer perceptron, we rewrite the above equation. Each feature map $y_i^{(l)}$ in layer l consists of $m_2^{(l)}.m_3^{(l)}$unit arranged in a two-dimensional array. The unit at position (r, s) computes the output

$$(y_i^{(i)})_{r,s} = \sum_{j=1}^{m1^{(l=1)}} (K_{i,j}^{(l)} * y_j^{(l-1)})_{r,s} + (b_i^{(l)})_{r,s} \quad (3)$$

$$= \sum_{j=1}^{m_1^{(l-1)}} \sum_{u=-h_1^{(l)}}^{h_1^{(l)}} \sum_{v=-h_2^{(l)}}^{h_2^{(l)}} \left(K_{i,j}^{(l)}\right)_{u,v} \left(y_j^{(l-1)}\right)_{r+u,s+v} + \left(b_i^{(l)}\right) \quad (4)$$

The trainable weight of the network can be found in the filters $K_{i,j}^{(l)}$ and bias matrices $b_i^{(l)}$.

2.5.2. Pooling layer

The pooling layer is to progressively reduce the spatial size of the representation to reduce the number of parameters and computation in the network, and hence to also control overfitting. It operates independently on every depth slice of the input and resizes it spatially, using the MAX operation. There are two types of pooling functions such as average pooling, and max pooling. Max-Pooling extracts the most important features like edges whereas, average pooling extracts features so smoothly. A max-pooling is better for extracting the extreme features while average pooling sometimes can't extract good features (it takes all into the count and results in an average value which may/may not be important for object detection type tasks). The pooling equation is given below:

$$S_p^1\ (i, j) = \frac{1}{4} \sum_{u=0}^{1} \sum_{v=0}^{1} y_p^1(2i - u,\ 2j - v),\ \ i, j = 1,\ 2 \ldots\ldots\ldots n.$$

2.5.3. Fully connected layer

If layer $l - 1$ is a fully connected layer which has been illustrated in equation (b). The layer l expects $m_1^{(l-1)}$ feature maps of size $m_2^{(l-1)} \times m_3^{(l-1)}$ as input and the i^{th} unit in layer l calculates.

$$y_i^{(l)} = f(c_i^{(l)} \text{ with } c_i^{(l)} = \sum_{j=1}^{m_1^{(l-1)}} \sum_{r=1}^{m_2^{(l-1)}} \sum_{s=1}^{m_3^{(l-1)}} w_{i,j,r,s}^{(l)} (y_j^{(l-1)})_{r,s}$$

2.6. Deep belief networks

Deep Neural Networks are neural networks that have relatively high depth.

The energy of a Boltzmann machine can be defined as

$$E(x, h) = \sum_{i \epsilon visible} a_i x_i - \sum_{j \epsilon\ hidden} b_j h_j - \sum_{i,j} x_i h_j w_{i,j} - \sum_{i,j} x_i x_j u_{i,j} - \sum_{i,j} h_i h_j v_{i,j}$$

Where v_i and h_j are the binary states of the visible unit i and hidden unit j, a_i and b_j are their biases, and $w_{i,j}$ is the weight connection between inputs.

Change in weight in an RBM is given by the learning rule:

$$\Delta w_{i,j} = \in (< v_i h_j >_{data} - \ < v_i h_j >_{model})$$

Where ε is the learning rate. Although, $< v_i h_j >_{data}$ is very easy but $< v_i h_j >_{model\, i}$ is very difficult to compute and a random selection could be considered in training vector v and then the binary state h_j of each of the hidden units is 1 with probability:

$$p\left(h_j = 1 | v\right) = \sigma(b_j + \sum_i v_i w_{i,j})$$

Where $\sigma(x)$ a logistic sigmoid function such as $1/(1 + \exp(-x))$. Mathematically, it can be described with l layers according to the joint distribution-

$$P\left(x,\ h^1, \ldots.,\ h^l\right) = \left(\prod_{k=0}^{l-2} P(h^k | h^{k+1})\right) P\left(h^{l-1}, h^l\right)$$

Where $x = h^0$, while $P(h^{k-1}|h^k)$ is a visible-given hidden conditional distribution in the RBM at level k of the DBN, and $P(h^{l-1}, h^1)$ is the top-level RBM's joint distribution.

3. Methods

3.1. Research design

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA), which is based on the Cochrane's Handbook for Systematic Reviews, was used to conduct the current study [26]. A review of the written protocol was not drafted (Supplementary Table S1). The process of study is given below (Fig. 5).

3.2. Search methods for identification of studies

3.2.1. Electronic database search

We systematically search in the widely used search engines such as EMBASE, PubMed, Scopus, Google Scholar, and Web of Science to find out relevant published between 1 January 2000 and 1 March 2019 using most appropriate free text terms ("Retinopathy" OR "Diabetic Retinopathy", OR "Referable diabetic retinopathy", OR "Retinal fundus image", OR "Vision threating diabetic retinopathy", OR "DR"), AND ("Deep learning" OR "DL" OR "Convolutional neural network" OR "CNN", "Deep neural network", OR "Automated technique", OR "Artificial intelligent").

3.2.2. Searching for other sources

We also carefully searched the reference lists of retrieved studies and relevant previous review articles for further inclusion.

3.3. Inclusion and exclusion criteria

Two authors (MMI, TNP) independently screened the titles and abstracts for retrieved articles primarily considered in the systematic review. They selected the most relevant and potential articles from the initial screening and kept it for full-text review. Disagreements were resolved by discussion with another expert (HCY). Finally, all discrepancies were settled by a discussion with the main investigator (YCL). To be included, all the studies had to fulfill the following criteria: 1) study must be in English and be peer-reviewed, 2) provided an outcome of DL algorithms and DR detection, 3) had to provide information any of the evaluation matric such as accuracy, AUROC, sensitivity and specificity, 4) had to provide clear information about database and number of images 5) had to provide a clear definition of DR, 6) clearly described DL algorithms and process used in the DR detection.

Editorials, short reports, traditional methods for detecting DR were excluded. All the studies meeting inclusion criteria at this stage were additionally reviewed by the same two authors to ensure the appropriateness of the final analysis. All disagreement between two authors for selecting potential studies were then resolved by the main investigator (YCL). Studies providing the most detailed information regarding algorithms were kept for references.

3.4. Data extraction

Two of us (MMI) and (TNP) individually retrieved all information from the selected articles based on the predefined, standard-

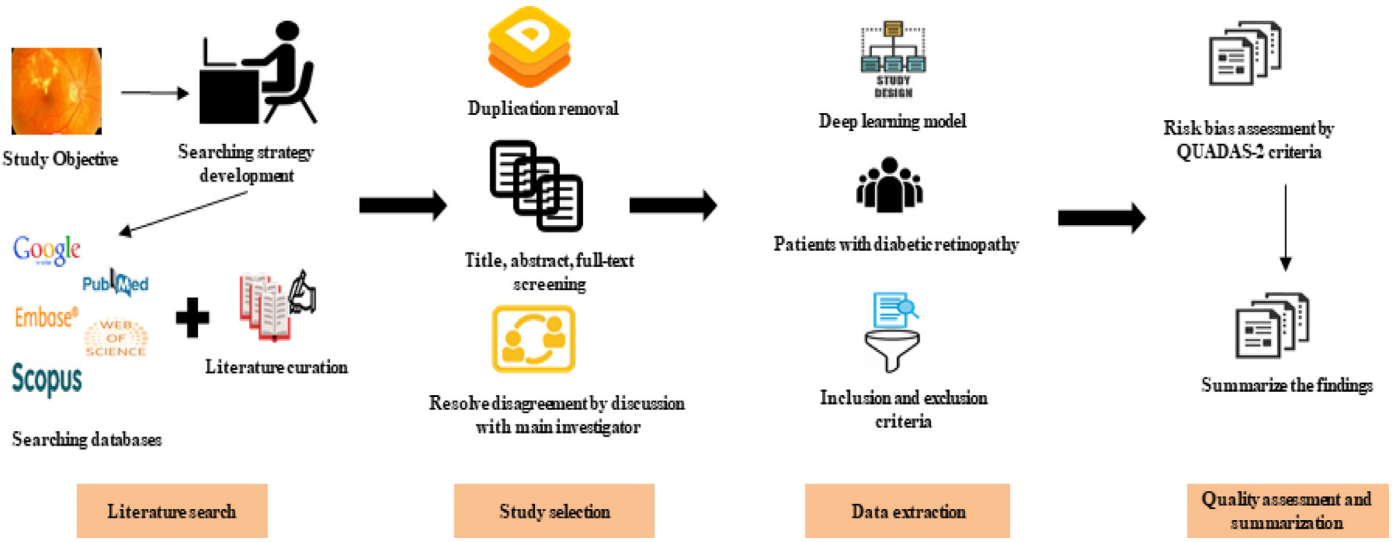


Fig. 5. Flow-chart of the Systematic Review Process.

ized protocol and data collection instrument. They entered the data to Review Manager software (RevMan-5) and subsequently checked for accuracy. Finally, they collected the following information from the included articles: (a) methods: model used in the study, database information, number of images, grade of retinopathy, identification process of retinopathy, inclusion and exclusion criteria, camera information, proportion of training and testing dataset, dataset used in training, testing and validation process, number of diabetic patients, evaluation metrics such as accuracy, sensitivity, specificity and AUROC; (b) participants: total number of participants, number of DR images, mean age, age range, gender, percentage of gender; (c) interventions: CNN, deep neural network; (d) outcome: prediction of DR (Referral, vision threating, and other types of DR).

3.5. Quality assessment

Any systematic reviews with meta-analysis of diagnostic accuracy studies are often characterized by some sorts of heterogeneous findings that originates from differences in the design and conduct of included studies [27]. MMI and TNP independently used the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool for evaluating the quality of the included studies. The QUADAS-2 scale [28] consists of four key domains for risk of bias such as patient selection, index test, reference standard, and flow and timing and three domains for applicability concerns such as patient selection, index test, and reference standard. The risk of bias was classified into three parts (low, high and unclear risk bias).

3.6. Outcome parameters

The two primary outcome parameters of this systematic review and meta-analysis were: (1) to evaluate the performance of DL algorithms for referable DR detection from the digital fundus images (2) to identify the source of bias and variation in diagnostic accuracy studies by QUADAS-2.

3.7. Statistical analysis

We used Meta-Disc (Version: 1.4) for statistical analysis such as a pooled estimate of AUROC, sensitivity, specificity and diagnostic odds ratio. It was used to a) summarize data from each individual study, b) evaluate the homogeneity of included studies graphically and statistically, c) calculate the pooled estimate. Diagnostic odds ratio (DOR) was also computed to know, how much greater the odds of having DR are for the people with a positive test result than for the people with a negative test result. DOR calculated mathematically in the following way,

$$DOR = LR + /LR - .$$

Likelihood ratios are calculated to express how much more frequent the respecting finding is among the individuals with DR than among the individuals without DR.

$$LR+ = (Sensitivity/1 - Specificity) \text{ and}$$

$$LR- = (1 - Sensitivity/Specificity)$$

The pooled AUROC of DL model, ≥90, <0.90, <80, <70, <0.60 defines as excellent, good, fair, poor and fail. An I^2 value was calculated to measure the statistical heterogeneity which reported an estimate of the percentage of variability among the included studies. An I^2 value of 0~25%, 25~50%, 50~75%, >75% represents very low, low, medium and high heterogeneity [27]. The value of I^2 was computed as follows:

$$I^2 = 100\% \ (Q - d.f)/Q$$

Here, Q = Cochrane's heterogeneity statistic and df = degree of freedom. Negative values of I^2 is considered as zero; the I^2 value is between 0% (no observed heterogeneity) and 100% (maximum heterogeneity).

4. Results

4.1. Study selection

The article search of the electronic database yielded 425 articles. A total of 322 articles were excluded due to duplication. After reviewing of all titles and abstracts, 70 articles were further excluded for lack of adherence to our study inclusion criteria mentioned in the method parts. 31 selected articles went for the full-text revision and also checked their reference lists for relevant articles, retrieving 2 additional articles. Afterward, 10 articles were excluded based on exclusion criteria (4 articles were excluded for

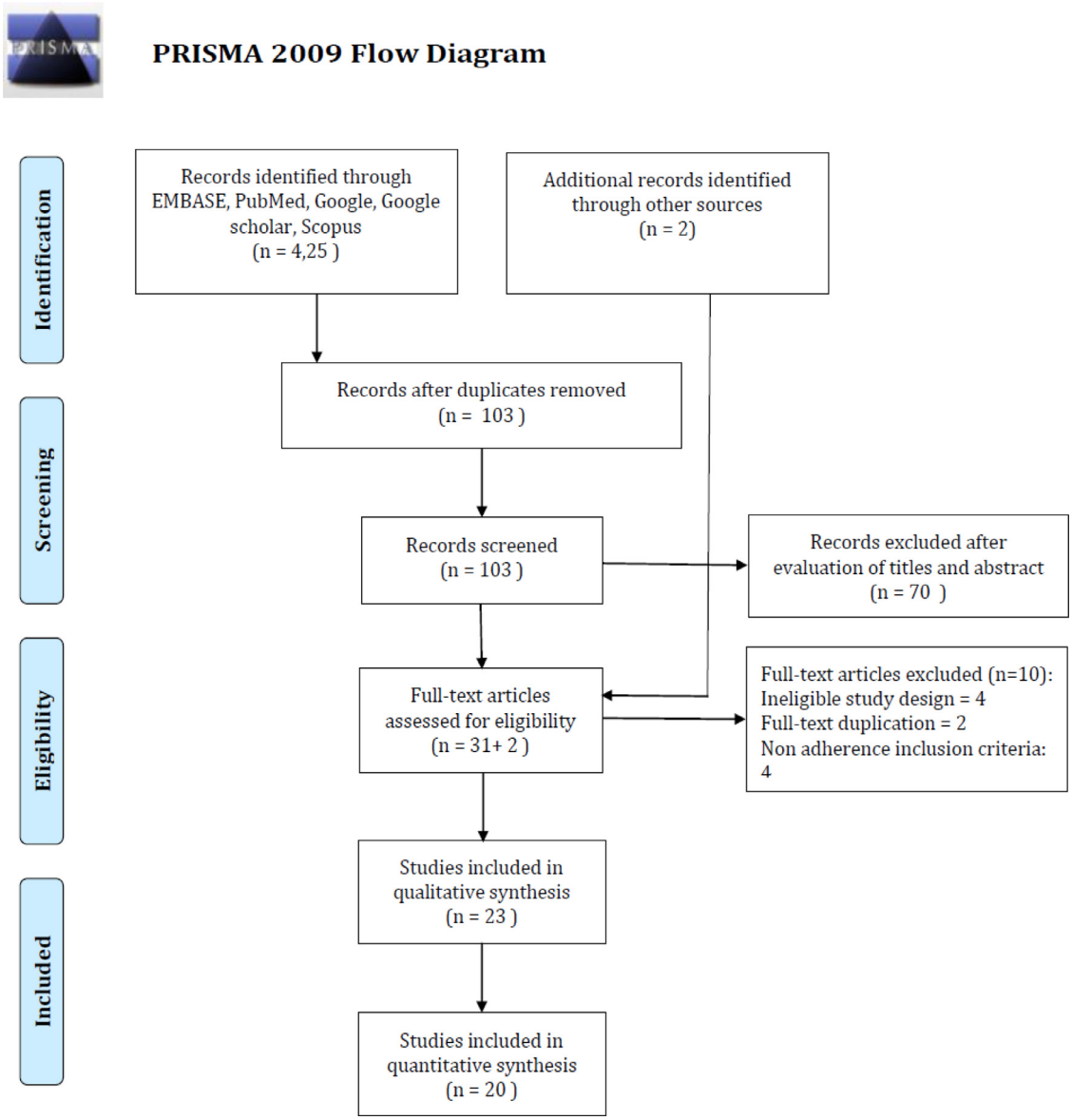


Fig. 6. PRISMA Flow Diagram for Study Selection.

non-adherence with the inclusion criteria, 4 articles were excluded for inappropriate study design, 2 studies excluded for duplication). Finally, 23 studies were selected for this systematic review [29-52]. 20 out of 23 studies met all inclusion criteria for meta-analysis. The flow diagram of the systematic literature review is presented in Fig. 6.

4.2. Study characteristics

The studies included in our systematic review with meta-analysis, comprised of 23 DL algorithms studies (Table 2). Publication years ranged from 2016 [29] to 2019 [32] and every study followed international clinical DR disease severity scale to grade DR images. All the included studies reported reference standards except one study [47]. Twenty studies outcome was DR or referable DR, and three studies showed DL performance on vision threating DR detection. Eighteen studies evaluated the performance of CNN for detecting DR from the retinal fundus images, three studies assessed the performance of hybrid CNN methods (CNN with random forest, decision tree and support vector machine) [47,32,35], two studies evaluated the performance of DCNN [43,37] and one study used third-party DL algorithm [33]. The range of sensitivity and specificity of included studies was 30 to 100 and 70.7 to 98.5. The range of AUROC was 95 to 99.3. Moreover, the accuracy of the DL model to detect/classify DR was 75 to 97.28. Most famous datasets such as EyePACS, SIDRP, SAMS and SPPH, Messidor, Messidor-2, Kaggle competition database were used to train, testing and validate the performance of DL algorithms (Table 3). 6 out of 23 studies used a different dataset evaluating the performance of their existing model.

4.3. DL and referable DR detection

Twenty studies evaluated the performance of a DL algorithm for automated detection of referable DR in retinal fundus photographs. The pooled AUROC for DL to detect rDR was 0.97 (95%CI: 0.95–0.98). The sensitivity and specificity of DL for rDR was 0.83 (95%CI: 0.83–0.83), and 0.92 (95%CI: 0.92–0.92), respectively. The positive-

Table 2
Characteristics of included studies.

Author	Year	Reference standard	Grading Scale	TD	VD/ TED	Model	Target	ACC (%)	SN (%)	SP (%)	AUROC (%)	External validation
Bellemo	2019	Two professional senior graders using ICDRSS	ICDRSS	SIDRP (n = 76,370 images)	Urban centers in the Copperbelt province of Zambia (4504 images from 1574 patients with DR	CNN	Rdr		92.25	89.04	97.3	N/A
Sayres	2019	Ophthalmologists, Retinal specialists	ICDRSS	EyePACS (n = 140,000 images)	EyePACS-2 (n = 1958)	CNN	rDR	88.4	91.55	94.69	–	N/A
Zeng	2019	Kaggle grade	ICDRSS	EyePACS (n = 28,104)	EyePACS (n = 7024)	CNN	rDR	–	82.2	70.7	95.1	N/A
Zhang	2019	ETDRS	GSDR	SAMS and SPPH (n = 3062 images)	SAMS and SPPH (3833 images)	CNN	rDR		97.5	97.7	97.7	Yes
Torre	2019	Standard severity scale	ICDRSS	EyePACS dataset (n = 75,650 images)	EyePACS dataset (3000 images as testing and 10,000 images as validation)	CNN	rDR	–	91.1	90.8	–	N/A
Pires	2019	ICDR & NPDR	ICDRSS	Kaggle competition dataset Messidor-2 dataset	Cross-dataset validation	CNN	rDR	–	95.8	83.3%	98.2	Yes
Li	2019	Ophthalmologists	ICDRSS	Kaggle competition-Dataset(n = 34,124 images)	Kaggle competition Dataset (1000 images for validation and 53,572 images for testing)	DCNN	DR	86.17	–	–	–	N/A
Li	2018	Ophthalmologists	ICDRSS	LabelMe (n = 58,792 images)	LabelMe (n = 8000)	CNN	vtDR	–	97	91.4	98.9	N/A
Seth	2018	NR	NR	EyePACS(n = 28,100 images)	EyePACS (n = 7026)	CNN with SVM	DR	–	93	85	–	N/A
Lin	2018	Two professional clinician	ICDRSS	Kaggle DR (n = 30,000 images were randomly selected)	Kaggle DR (3000 images were used as the testing set)	CNN		86.10	73.24	93.81		N/A
Keel	2018	Human grader at a centralized retinal grading center	DR: R0-R3 DME: M0, M1, P, U.	Online dataset from China (n = 66,790 image)	96 patients from 2 outpatient clinics in Australia (n = 192)	Third-party DL algorithm	rDR	–	92.3	93.7	99.3	N/A

(*continued on next page*)

Table 2 (*continued*)

Author	Year	Reference standard	Grading Scale	TD	VD/ TED	Model	Target	ACC (%)	SN (%)	SP (%)	AUROC (%)	External validation
Chang	2018	Licensed clinician (Kaggle label)	DR: 0–4 (DME not included)	Kaggle DR competition dataset (n = 35,126 images)	Kaggle DR competition dataset (n = 35,126)	CNN	DR	78.7	–	–	–	N/A
Wan	2018	Medical Experts	ICDRSS	Kaggle DR	Kaggle DR	CNN	rDR	95.68	86.47	97.43	97	N/A
Gargeya	2017	Medical experts (Messidor-2 label)	DR: R0–R3 DME: 0–2	EyePACS (n = 75,125 images)	Messidor-2 (n = 1748) & E-Ophtha (n = 463)	CNN	DR	–	93 & 90	87 & 95	94 & 95	Yes
Garcia	2017	Medical expert	ICDRSS	EyePACS	EyePACS	CNN	DR	83.68	54.47	93.65	–	N/A
Takahashi	2017	Human grader	DR: modified Davis grading (0–3).	Japanese hospital data (n = 9443)	Japanese hospital (n = 496)	CNN	DR	96	–	–	–	N/A
Ting	2017	Ophthalmologist, Retinal specialist	ICDRSS DME: 0–1	SIDRP: Year: 2010–2013) (n = 76,370 images)	SIDRP: Year: 2014–2015) (n = 71,752) & 10 others external dataset (n = 40,752)	CNN	rDR &vtDR	–	90.5 & 100	91.6 & 91.1	93.6 & 95.8	Yes
Mansour	2017	Licensed clinicians	ICDRSS	Image Net (n = 14,197,122 images)	Kaggle dataset (n = 35,126)	CNN	DR	97.28	100	99	–	N/A
Ramachandran	2017	Ophthalmologist	Otago: 6 class and Messidor: 4 class	ODEMS	ODEMS & Messidor	DNN	rDR	–	84.6 & 96	79.7 & 90	90.1 & 98	Yes
Raju	2017	Ophthalmologists	ICDRSS	EyePACS (n = 35,126 images)	Kaggle data (n = 53,126)	CNN	DR	93.28	80.28	92.29	–	N/A
Abràmoff	2016	Consensus among 3 US board certified retinal specialists	DR: ICDR DME:modified definition of DME (0–2)	EyeCheck project and Iowa University (10,000 – 1250,000 unique sample extracted from images from patients with DR	Messidor-2 (874 subjects with diabetics, 1748 images)	A hybrid model with multiple CNN and Random forest classifier	rDR vtDR DME	–	96.8 (93.3–98.8)	87 (84.2–89.4)	98 (96–99)	N/A
Gulshan	2016	US certified ophthalmologists	ICDRSS	EyePACS (n = 94,281 images) & 3 eye hospital in India (n = 33,894 images)	Messidor-2 (n = 1748) & EyePACS-1 (n = 9963)	CNN	rDR	–	87 & 90.3	98.5 & 98.1	99 & 99.1	Yes
Pratt	2016	Kaggle grades	ICDRSS	Kaggle dataset (n = 80,000 images)	Kaggle dataset (n = 5000)	CNN	DR	75	30	95	–	N/A

Note: SAMS and SPPH = the Sichuan Academy of Medical Sciences and Sichuan Provincial Peoples Hospital, CNN= Convolutional Neural Network, DCNN= Deep Convolutional Neural Network, DR= Diabetic retinopathy, rDR= referable diabetic retinopathy, vtDR= vision-threatening diabetic retinopathy, DME: diabetic macular edema, ICDRSS: International Clinical Diabetic Retinopathy disease severity scale, SIDRP= Singapore National Diabetes Retinopathy Screening Program, ACC= Accuracy, SN= Sensitivity, SP= Specificity, AUROC= Area Under the Receiver Operating Curve, DL= Deep learning, SVM= Support Vector Machine, US= United States, ODEMS= Recurrent optic disc edema with a macular star, GSDR= Grading Scale of Diabetic Retinopathy, ETDRS= The Early Treatment in Diabetic Retinopathy Study, NPDR=Mild Non-proliferative Retinopathy, TD= Training dataset, VD= Validation dataset, N/*A*= Not applicable.

Table 3
Description of Databases.

Dataset	Number of images	Description	Camera	Pupil dilation
EyePACS	35,126	It is a non-proprietary, freely accessible and open source web-based application for exchanging eye-related clinical information in the USA	A variety of digital cameras including Centervue DRS, Opovue iCam, Canon CR1/DGi/CR2 and Topcon NW using 45° fields of view.	40%
EyePACS-1	9963	Images were randomly taken from original EyePACS dataset	A variety of digital cameras including Centervue DRS, Opovue iCam, Canon CR1/DGi/CR2 and Topcon NW using 450 fields of view.	40%
E-Ophtha	107,799	OPHDIAT Tele-medical network for DR screening in France. There is a variation in the size and resolution of the images, ranging from 1440×960 to 2544×1696 pixels. All images were resized to the size of the DIARETDB1 (1500×1152 pixels).	The digital color fundus camera	NR
Kaggle DR competition dataset	88,702	Primary care sites in California and elsewhere in the US. Images are uploaded to the EyePACS DR screening platform.	Digital retinal camera	NR
Messidor	1200	Consecutive enrolment at three ophthalmological departments in France. Images were captured using 8 bits per color plane at 1440×960, 2240×1488 or 2304×1536 pixels.	A color video 3CCD camera mounted on a Topcon TRC NW6 non-mydriatic retinal-graph with a 45-degree field of view.	800 images were acquired with pupil dilation (one drop of Tropicamide at 0.5%) and 400 without dilation
Messidor-2	1748	Consecutive enrolment at three ophthalmological departments in France. Images were captured using 8 bits per color plane at 1440×960, 2240×1488 or 2304×1536 pixels	A Topcon TRC NW6 nonmydriatic camera and 45° fields of view centered on the fovea.	Approximately 44%
Otago	≥1764	Local diabetic retinal screening program in New Zealand	Digital non-mydriatic retinal camera with 45° fields of view	75%
SIDRP 2014–2015	71,896	Singapore National DR Screening Program from Singapore, Malaysia	Digital camera including Topcon, Canon, Carl Zeiss camera centered at optic disc and fovea	NR
Royal Victoria Eye and Ear Hospital	2302	Clinic-based Australian Hospital	Different digital camera	NR
Australian patients from outpatient clinics	192	Two Australian endocrinological outpatient clinics.	A digital camera with 45° fields of view pointed at a central-nasal field	NR
Patients from a Japanese hospital	9939	Medical university in Japan.	Non-mydriatic color fundus digital camera with 45° fields of view.	NR
SAMS and SPPH	13,767	The Sichuan Academy of Medical Sciences and Sichuan Provincial Peoples Hospital	Photographs with a 45° view	NR

NR = Not reported.

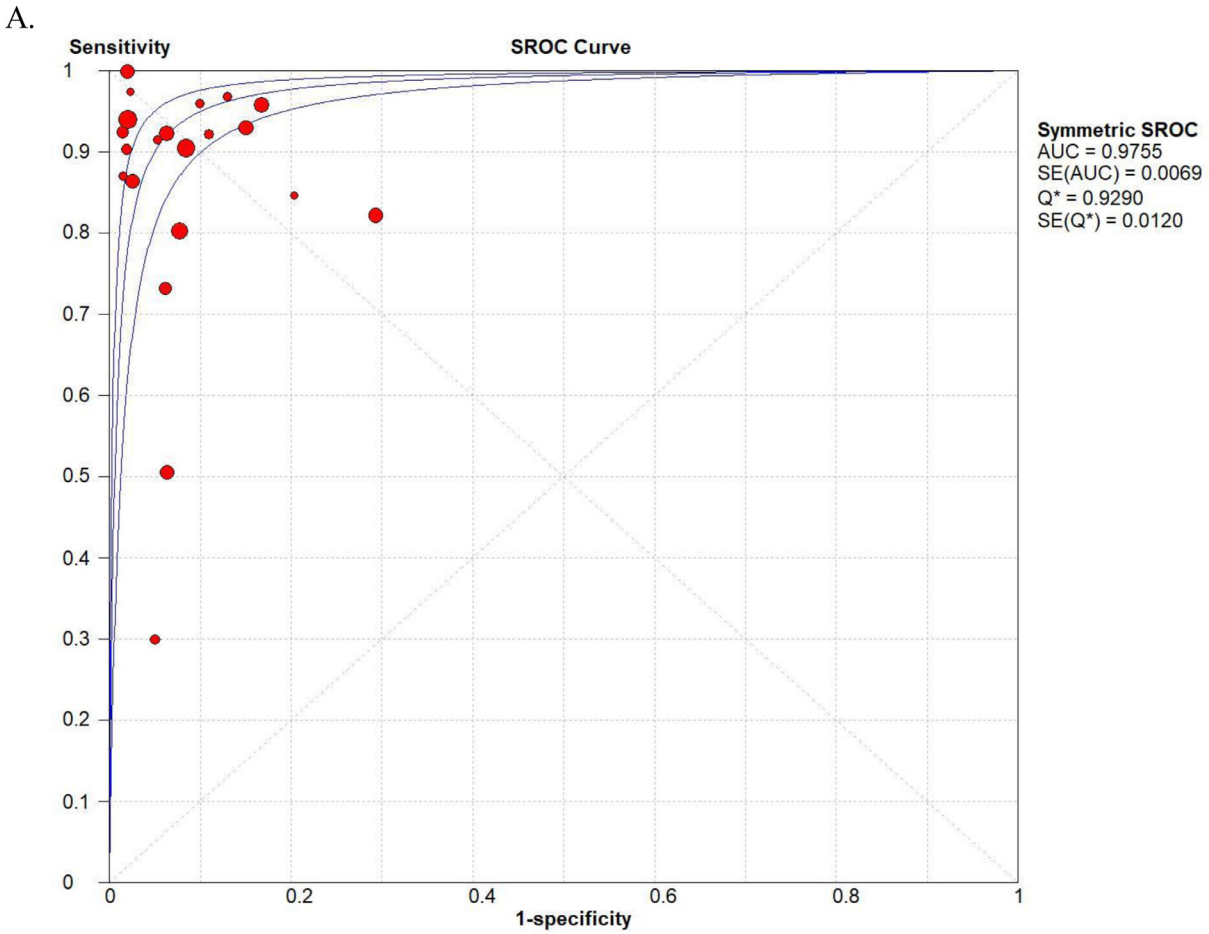


Fig 7. Performance of the DL model for detecting rDR (A. AUROC, B. Sensitivity C. Specificity D. Diagnostic odds ratio). The color circle represents the proportion of the number of patients with DR and the number of patients without DR. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and negative-likelihood ratio was 14.11 (95%CI: 9.91–20.07), and 0.10 (95%CI: 0.07–0.16). Moreover, the diagnostic odds ratio was 136.83 (95%CI: 79.03–236.93) (Fig. 7).

4.4. Subgroup analysis

Two studies evaluated the performance of the DL model for the detection of vision-threatening DR. The sensitivity and specificity of the DL for vision-threating DR was 0.92 (95%CI:0.90–0.94), and 0.91 (95%CI: 0.90–0.92). The positive likelihood ratio, negative likelihood ratio, and diagnostic odds ratio were 11. 97 (95%CI: 5.91-24.23), 0.09 (95%CI: 0.07–0.11) and 131.89 (95%CI: 66.94–259.87), respectively.

4.5. Risk of bias and applicability

In this current study, we also assessed heterogeneous findings that originated from included studies based on the QUADAS-2 tool (Table 4). The risk of bias for patient's selection was unclear for twelve studies. Five studies had an unclear risk of bias for flow and timing and only one study had an unclear risk of bias for the reference standard. Moreover, four studies had a higher risk of bias for the reference standard and one study had a higher risk of bias patient selection and one study had a higher risk of bias for flow and timing. On the other hand, six studies had a higher risk of applicability concern for the reference standard.

5. Discussion

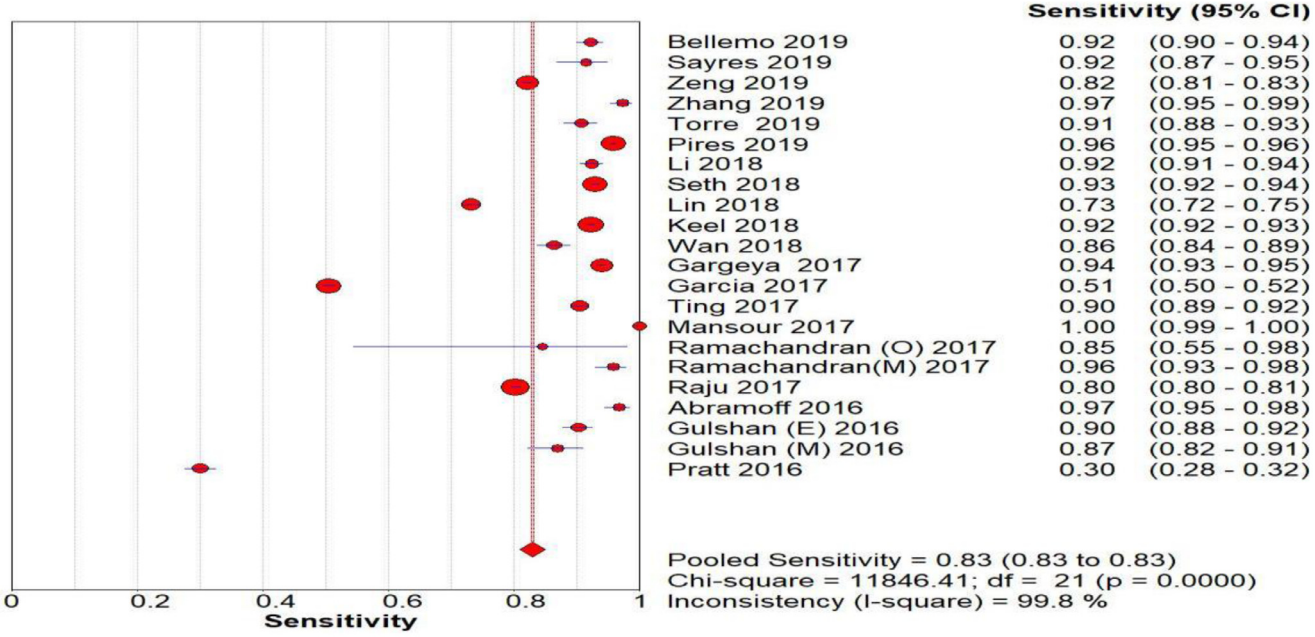
5.1. Main findings

This systematic review with meta-analysis evaluated the performance of DL algorithms to detect referable DR automatically using color fundus retinal images as compared to the standard method available in clinical settings. The key findings are (a) DL algorithms (DCNN/CNN) achieved high sensitivity and specificity in detecting referable DR from retinal fundus images, (b) features analysis of color fundus images using CNN provided insights about the diagnostic process used by retinal specialists. There are several techniques that can accurately detect DR but it is labor-intensive and required expert ophthalmologists. They also can be time-consuming and uncomfortable for patients with DR due to higher data acquisition and interpretation time. A DL-based automated screening tool could offer an alternative solution for DR screening, especially in settings with little access to human expertise.

5.2. Public health implication

A significant proportion of patients with diabetes develop DR that usually has no symptoms until the very last stages [53]. It still remains the main reason for effective treatments and often leads to the blindness of these patients. Furthermore, clinical diagnosis

B.



C.

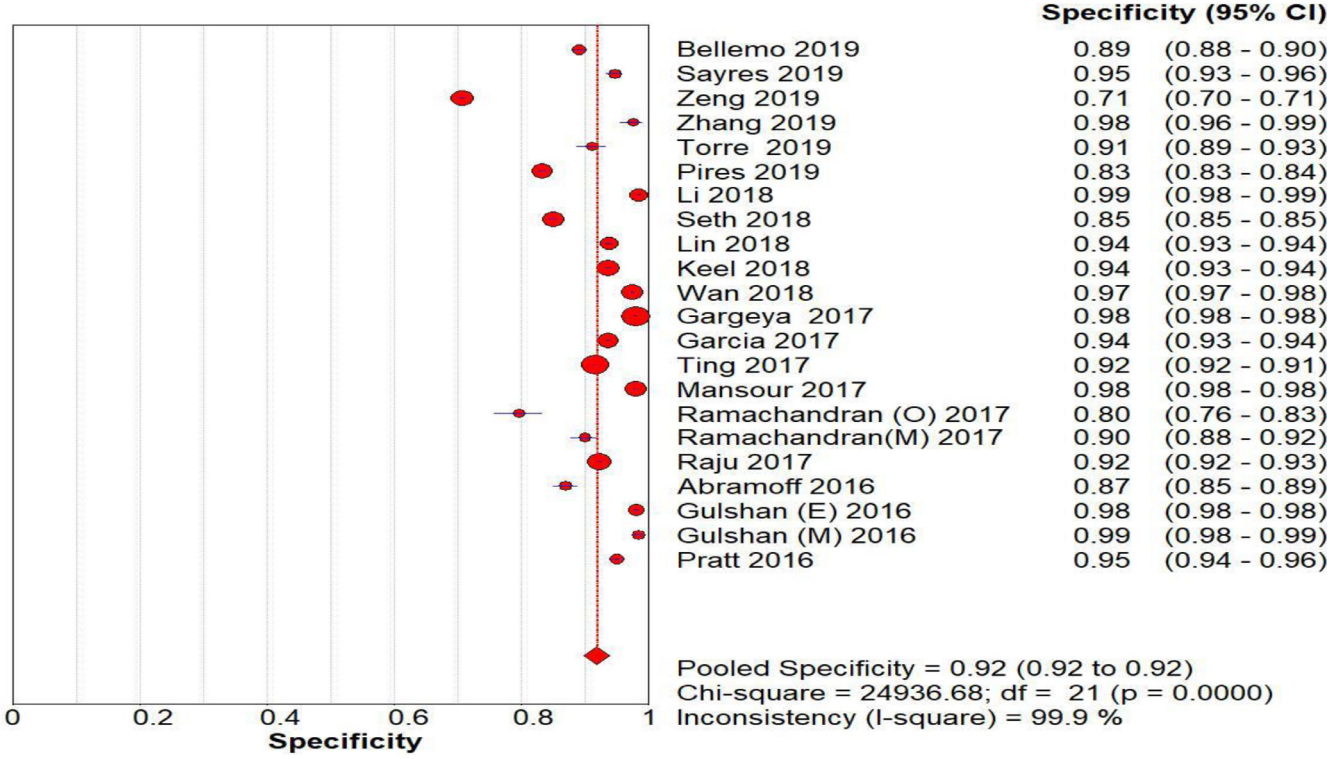


Fig 7. Continued

of DR is solely based on the appearance of retinal vessels but all the process is highly subjective and qualitative. Despite developing the revised guidelines to mitigate DR, the prevalence of vision-threatening DR is high [54]. Primary challenges of detecting DR are (a) highly variable and expert ophthalmologist dependent (b) scarcity of retinal experts, and willingness to manage DR is insufficient. It is mainly because of infrastructure, lack of training process, time-consuming examination process, public awareness, and significant malpractice liability [55-57], and (c) an increasing number of diabetic patients. These challenges have triggered the need to develop quantitative and objective approaches to DR diagnosis using DL-based automatic detection. Multiple studies have already

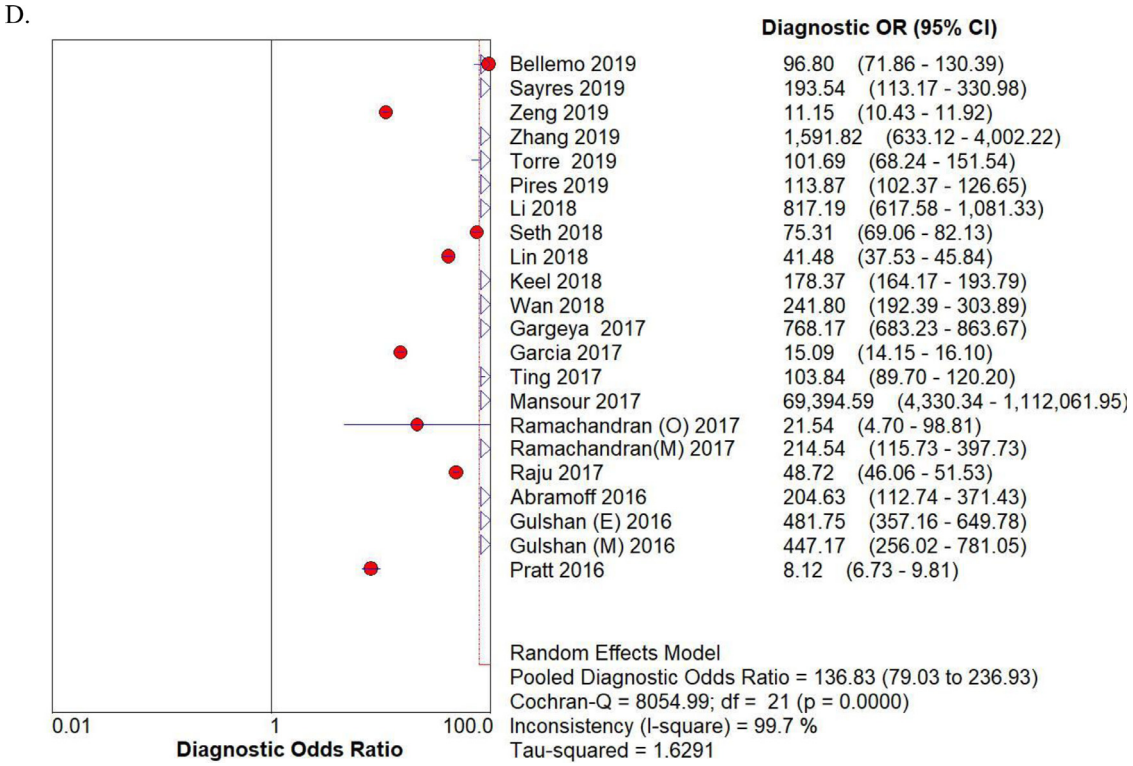


Fig 7. Continued

shown that DL algorithms outperformed clinician performance for detecting DR [29,31]. Furthermore, this current systematic review with meta-analysis also showed that DL achieved tremendous performance in detecting DR; therefore, a fully automated tool based on DL would improve the quality of care, and enhance the accessibility of care by making effective and efficient automated screening systems.

5.3. Opportunities and challenges

There are immense opportunities for using DL algorithms for the automated identification of DR from the color fundus images. Included studies in this systematic review showed a promising performance to perform a screening function that has clinical relevance with the physician's screening performance. Implementation of DL-based automated systems would (a) improve the efficiency and coverage (As the manual identification depends on experience, skill and other factors but algorithms could do it repetitively and performance would be consistency), (b) reduce barriers to access the remote areas where skilled ophthalmologists are not always accessible, (c) reduce the need of manpower, (d) help to detect DR at a stage before converting to PDR (treatment of DR at advance stages with conventional laser therapy, intravitreal anti-VEGF or corticosteroid injection are expensive and create serious complication for patients), (e) decrease healthcare cost through earlier intervention in treatable stage. There are also several challenges that need to be addressed. First, the way DL model works, especially CNN provided with only the image and associated grade, without explicit definitions of features (e.g. micro-aneurysms, exudates). Second, a generalization of the DL model, which means how good or bad the model performs on new data. It is very important for the DL model to be generalized for building trust and be able to rely on a DL-based automated system. Third, the black-box issues that could affect ophthalmologists acceptance the performance of the DL algorithms for clinical use. Fourth, different levels of DR (e.g. PDR, ME) from the fundus images may not be correctly classified all cases appropriately without clinical examination because it may be affected by other eye diseases. Fifth, the performance of the CNN model is based on the quality of images on which the CNN model is making decisions and diagnosis (may not perform low-quality images or non-fundus images). Finally, the CNN model could not identify non-DR lesion (non-graded) because most of the cases, CNN models are trained to identify only DR and DME.

5.4. Strengths and limitations

Our study has several strengths and limitations. First, this the most comprehensive systematic review and meta-analysis that evaluated the performance of the DL model to detect DR. A total of 23 studies were included with detail information about DL algorithms and the risk of bias and applicability concerns of included studies were also evaluated. Second, included studies used different kinds of the dataset including EyePACs which is large and widely accepted standard image dataset for DR. Our study has several limitations. First, although we included 20 large studies in our meta-analysis; but heterogeneity among the study was high. Second, we are unable to assume how the DL algorithms like the CNN model would perform in real-world DR screening program. Finally, most of the included studies used a common reference standard for image classification decisions taken by ophthalmologist graders. This means the algorithm may not perform well for images with subtle findings that a majority of ophthalmologists would not identify.

Table 4
Quality Assessment of Diagnostic Accuracy Studies-2 for Included Studies.

Study	Risk of Bias				Applicability Concerns		
	Patients Selection	Index Test	Reference Standard	Flow and Timing	Patients Selection	Index Test	Reference Standard
Bellemo 2019	☺	☺	☺	☺	☺	☺	☺
Sayres 2019	☺	☺	☺	☺	?	☺	☺
Zeng 2019	?	☺		☺	?	☺	
Zhang 2019	☺	☺		☺	?	☺	☺
Torre 2019	☺	☺	☺	☺	?	☺	☺
Pires 2019	?	☺	☺	☺	?	☺	☺
Li 2019	☺	☺	☺	?	?	☺	☺
Li 2019	?	☺	☺		?	☺	?
Seth 2018	?	☺		?	?	☺	
Lin 2018	?	☺	☺	?	?	☺	☺
Keel 2018	?	☺	☺	?	☺	☺	☺
Chang 2018		☺	☺	☺	?	☺	
Mansour 2018	?	☺	☺	☺	?	☺	
Ramachandran 20018	?	☺		☺	?	☺	
Wan 2018	☺	☺	☺	?	?	☺	☺
Gargeya 2017	☺	☺	☺	☺	?	☺	☺
Garcia 2017	?	☺	☺	☺	?	☺	?
Takahashi 2017	?	☺	☺	☺	?	☺	☺
Ting 2017	?	☺	☺	☺	☺	☺	☺
Raju 2017	?	☺	☺	☺	?	☺	
Abràmoff 2016	☺	☺	☺	☺	?	☺	☺
Gulshan 2016	☺	☺	☺	☺	?	☺	☺
Pratt 2016	☺	☺	?	☺	?	☺	?

Note: ☺ = No risk bias; ? = Unclear risk bias; = High risk of bias

6. Conclusion

he findings of our study show that DL can play an important role in detecting referable DR with high sensitivity, specificity, and repeatability. The application of a DL-based automated system may change the way DR is diagnosed in the future. Automated tools may improve the quality of DR screening, accessibility to healthcare as well as to reduce the cost of screening. Earlier detection and timely treatment might prevent the onset of the disease or help to stop the progression at an earlier stage.

Declaration of Competing Interest

None.

Acknowledgement

This research a is sponsored in part by Taipei Medical University under grant TMU107-AE1-B18.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2020.105320.

References

[1] P. Zimmet, K.G. Alberti, D.J. Magliano, P.H. Bennett, Diabetes mellitus statistics on prevalence and mortality: facts and fallacies, Nat. Rev. Endocrinol. 12 (10) (2016) 616.
[2] T.N. Poly, M.M. Islam, H.C. Yang, P.-A. Nguyen, C.C. Wu, Y. Li, Artificial intelligence in diabetic retinopathy: insights from a meta-analysis of deep learning, Stud. Health Technol. Inform. 264 (2019) 1556–1557.
[3] J.L. Harding, M.E. Pavkov, D.J. Magliano, J.E. Shaw, E.W. Gregg, Global trends in diabetes complications: a review of current evidence, Diabetologia 62 (1) (2019) 3–16.
[4] L. Bäcklund, P. Algvere, U. Rosenqvist, New blindness in diabetes reduced by more than one-third in stockholm county, Diabetic Med. 14 (9) (1997) 732–740.
[5] R.A. Gangwani, J. Lian, S. McGhee, D. Wong, K.K. Li, Diabetic Retinopathy screening: Global and Local Perspective, Hong Kong Medical Journal, 2016.
[6] M.J. Fowler, Microvascular and macrovascular complications of diabetes, Clin. Diabetes 26 (2) (2008) 77–82.
[7] N.G. Congdon, D.S. Friedman, T. Lietman, Important causes of visual impairment in the world today, JAMA 290 (15) (2003) 2057–2060.
[8] Y.G. Park, Y.-J. Roh, New diagnostic and therapeutic approaches for preventing the progression of diabetic retinopathy, J. Diabetes Res. 2016 (2016).
[9] E. Stefánsson, T. Bek, M. Porta, N. Larsen, J.K. Kristinsson, E. Agardh, Screening and prevention of diabetic blindness, Acta Ophthalmol. Scand. 78 (4) (2000) 374–385.
[10] C.-H. Zhu, S.-S. Zhang, Y. Kong, Y.-F. Bi, L. Wang, Q. Zhang, Effects of intensive control of blood glucose and blood pressure on microvascular complications in patients with type ii diabetes mellitus, Int. J. Ophthalmol. 6 (2) (2013) 141.
[11] E.S. Moghissi, M.T. Korytkowski, M. DiNardo, D. Einhorn, R. Hellman, I.B. Hirsch, S.E. Inzucchi, F. Ismail-Beigi, M.S. Kirkman, G.E. Umpierrez, American association of clinical endocrinologists and American diabetes association consensus statement on inpatient glycemic control, Diabetes Care 32 (6) (2009) 1119–1131.
[12] S. Kuo, B.B. Fleming, N.S. Gittings, L.F. Han, L.S. Geiss, M.M. Engelgau, S.H. Roman, Trends in care practices and outcomes among Medicare beneficiaries with diabetes, Am. J. Prev. Med. 29 (5) (2005) 396–403.
[13] H. Lau, Y. Voo, K. Yeo, S. Ling, A. Jap, Mass screening for diabetic retinopathy–a report on diabetic retinal screening in primary care clinics in Singapore, Singapore Med. J. 36 (5) (1995) 510–513.
[14] P.A. Newcomb, R. Klein, Factors associated with compliance following diabetic eye screening, J. Diabet. Complications 4 (1) (1990) 8–14.
[15] T.Y. Wong, J. Sun, R. Kawasaki, P. Ruamviboonsuk, N. Gupta, V.C. Lansingh, M. Maia, W. Mathenge, S. Moreker, M.M. Muqit, Guidelines on diabetic eye care: the international council of ophthalmology recommendations for screening, follow-up, referral, and treatment based on resource settings, Ophthalmology 125 (10) (2018) 1608–1622.
[16] P. Burlina, D.E. Freund, B. Dupas, N. Bressler, Automatic screening of age-related macular degeneration and retinal abnormalities, in: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2011, pp. 3962–3966.
[17] A.K. Feeny, M. Tadarati, D.E. Freund, N.M. Bressler, P. Burlina, Automated segmentation of geographic atrophy of the retinal epithelium via random forests in AREDS color fundus images, Comput. Biol. Med. 65 (2015) 124–136.
[18] G. Gardner, D. Keating, T.H. Williamson, A.T. Elliott, Automatic detection of diabetic retinopathy using an artificial neural network: a screening tool, Br. J. Ophthalmol. 80 (11) (1996) 940–944.
[19] E.T.D.R.S.R. Group, Grading diabetic retinopathy from stereoscopic color fundus photographs—an extension of the modified Airlie House classification: ETDRS report number 10, Ophthalmology 98 (5) (1991) 786–806.
[20] C.J. Rudnisky, B.J. Hinz, M.T. Tennant, A.R. de Leon, M.D. Greve, High-resolution stereoscopic digital fundus photography versus contact lens biomicroscopy for the detection of clinically significant macular edema, Ophthalmology 109 (2) (2002) 267–274.
[21] S. Zachariah, W. Wykes, D. Yorston, Grading diabetic retinopathy (DR) using the Scottish grading protocol, Community Eye Health 28 (92) (2015) 72.
[22] F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain, Psychol. Rev. 65 (6) (1958) 386.
[23] M. Minsky, S.A. Papert, Perceptrons: An Introduction to Computational Geometry, MIT press, 2017.
[24] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436.
[25] A. Krizhevsky, I. Sutskever, G. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[26] D. Moher, A. Liberati, J. Tetzlaff, D.G. Altman, Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement, Ann. Intern. Med. 151 (4) (2009) 264–269.
[27] M.M. Islam, T. Nasrin, B.A. Walther, C.-C. Wu, H.-C. Yang, Y.-C. Li, Prediction of sepsis patients using machine learning approach: a meta-analysis, Comput. Methods Programs Biomed. (2019) 1–9 170.
[28] P.F. Whiting, A.W. Rutjes, M.E. Westwood, S. Mallett, J.J. Deeks, J.B. Reitsma, M.M. Leeflang, J.A. Sterne, P.M. Bossuyt, QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies, Ann. Intern. Med. 155 (8) (2011) 529–536.
[29] V. Gulshan, L. Peng, M. Coram, M.C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, JAMA 316 (22) (2016) 2402–2410.
[30] R. Gargeya, T. Leng, Automated identification of diabetic retinopathy using deep learning, Ophthalmology 124 (7) (2017) 962–969.
[31] D.S.W. Ting, C.Y.-L. Cheung, G. Lim, G.S.W. Tan, N.D. Quang, A. Gan, H. Hamzah, R. Garcia-Franco, I.Y. San Yeo, S.Y. Lee, Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes, JAMA 318 (22) (2017) 2211–2223.
[32] V. Bellemo, Z.W. Lim, G. Lim, Q.D. Nguyen, Y. Xie, M.Y. Yip, H. Hamzah, J. Ho, X.Q. Lee, W. Hsu, Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study, Lancet Digital Health 1 (1) (2019) e35–e44.
[33] S. Keel, P.Y. Lee, J. Scheetz, Z. Li, M.A. Kotowicz, R.J. MacIsaac, M. He, Feasibility and patient acceptability of a novel artificial intelligence-based screening model for diabetic retinopathy at endocrinology outpatient services: a pilot study, Sci. Rep. 8 (1) (2018) 4330.
[34] K. Chang, N. Balachandar, C. Lam, D. Yi, J. Brown, A. Beers, B. Rosen, D.L. Rubin, J. Kalpathy-Cramer, Distributed deep learning networks among institutions for medical imaging, J. Am. Med. Inf. Assoc. 25 (8) (2018) 945–954.
[35] M.D. Abràmoff, Y. Lou, A. Erginay, W. Clarida, R. Amelon, J.C. Folk, M. Niemeijer, Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning, Invest. Ophthalmol. Vis. Sci. 57 (13) (2016) 5200–5206.
[36] H. Takahashi, H. Tampo, Y. Arai, Y. Inoue, H. Kawashima, Applying artificial intelligence to disease staging: deep learning for improved staging of diabetic retinopathy, PLoS ONE 12 (6) (2017) e0179790.
[37] N. Ramachandran, S.C. Hong, M.J. Sime, G.A. Wilson, Diabetic retinopathy screening using deep neural network, Clin. Exp. Ophthalmol. 46 (4) (2018) 412–416.
[38] R.F. Mansour, Deep-learning-based automatic computer-aided diagnosis system for diabetic retinopathy, Biomed Eng. Lett. 8 (1) (2018) 41–57.
[39] M. Raju, V. Pagidimarri, R. Barreto, A. Kadam, V. Kasivajjala, Aswath a development of a deep learning algorithm for automatic diagnosis of diabetic retinopathy, in: MedInfo, 2017, pp. 559–563.
[40] G. García, J. Gallardo, A. Mauricio, J. López, C. Del Carpio, Detection of diabetic retinopathy based on a convolutional neural network using retinal fundus images, in: International Conference on Artificial Neural Networks, Springer, 2017, pp. 635–642.
[41] H. Pratt, F. Coenen, D.M. Broadbent, S.P. Harding, Y. Zheng, Convolutional neural networks for diabetic retinopathy, Procedia Comput. Sci. 90 (2016) 200–205.
[42] K. Xu, D. Feng, H. Mi, Deep convolutional neural network-based early automated detection of diabetic retinopathy using fundus image, Molecules 22 (12) (2017) 2054.
[43] Y.-H. Li, N.-N. Yeh, S.-J. Chen, Y.-C. Chung, Computer-assisted diagnosis for diabetic retinopathy based on fundus images using deep convolutional neural network, Mob. Inf. Syst. (2019) 2019.
[44] G.-M. Lin, M.-J. Chen, C.-H. Yeh, Y.-Y. Lin, H.-Y. Kuo, M.-H. Lin, M.-C. Chen, S.D. Lin, Y. Gao, A. Ran, Transforming retinal photographs to entropy images in deep learning to improve automated detection for diabetic retinopathy, J. Ophthalmol. (2018) 2018.

[45] R. Pires, S. Avila, J. Wainer, E. Valle, M.D. Abramoff, A. Rocha, A data-driven approach to referable diabetic retinopathy detection, Artif. Intell. Med. 96 (2019) 93–106.
[46] R. Sayres, A. Taly, E. Rahimy, K. Blumer, D. Coz, N. Hammel, J. Krause, A. Narayanaswamy, Z. Rastegar, D. Wu, Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy, Ophthalmology 126 (4) (2019) 552–564.
[47] S. Seth, B. Agarwal, A hybrid deep learning model for detecting diabetic retinopathy, J. Stat. Manage. Syst. 21 (4) (2018) 569–574.
[48] S. Wan, Y. Liang, Y. Zhang, Deep convolutional neural networks for diabetic retinopathy detection by image classification, Comput. Electr. Eng. 72 (2018) 274–282.
[49] X. Zeng, H. Chen, Y. Luo, W. Ye, Automated diabetic retinopathy detection based on binocular siamese-like convolutional neural network, IEEE Access 7 (2019) 30744–30753.
[50] W. Zhang, J. Zhong, S. Yang, Z. Gao, J. Hu, Y. Chen, Z. Yi, Automated identification and grading system of diabetic retinopathy using deep neural networks, Knowl. Based Syst. 175 (2019) 12–25.
[51] Z. Li, S. Keel, C. Liu, Y. He, W. Meng, J. Scheetz, P.Y. Lee, J. Shaw, D. Ting, T.Y. Wong, An automated grading system for detection of vision-threatening referable diabetic retinopathy on the basis of color fundus photographs, Diabetes Care 41 (12) (2018) 2509–2516.
[52] de La Torre J., Valls A., Puig D. (2019) A Deep Learning Interpretable Classifier For Diabetic Retinopathy Disease Grading. Neurocomputing
[53] C.E. Fraser, D.J. DAmico, Diabetic Retinopathy: Classification and Clinical Features, 2015 UpToDate (Waltham: UpToDate, 2014).
[54] J.M. Brown, J.P. Campbell, A. Beers, K. Chang, S. Ostmo, R.P. Chan, J. Dy, D. Erdogmus, S. Ioannidis, J. Kalpathy-Cramer, Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks, JAMA Ophthalmol. 136 (7) (2018) 803–810.
[55] D.K. Wallace, Fellowship training in retinopathy of prematurity, J. Am. Assoc. Pediatr. Ophthalmol. Strabismus 16 (1) (2012) 1.
[56] J.M. Furtado, V.C. Lansingh, K.L. Winthrop, B. Spivey, Training of an ophthalmologist in concepts and practice of community eye health, Indian J. Ophthalmol. 60 (5) (2012) 365.
[57] W.M. Fierson, A. Capone, O. AAoPSo, Telemedicine for evaluation of retinopathy of prematurity, Pediatrics 135 (1) (2015) e238–e254.