Review

Performance and Limitation of Machine Learning Algorithms for Diabetic Retinopathy Screening: Meta-analysis

Jo-Hsuan Wu[1], MD; T Y Alvin Liu[2], MD; Wan-Ting Hsu[3], MSc; Jennifer Hui-Chun Ho[4], PhD, MD; Chien-Chang Lee[5,6,7], SCD, MD

[1]Shiley Eye Institute and Viterbi Family Department of Ophthalmology, University of California San Diego, La Jolla, CA, United States

[2]Retina Division, Wilmer Eye Institute, The Johns Hopkins Medicine, Baltimore, MD, United States

[3]Harvard TH Chan School of Public Health, Boston, MA, United States

[4]National Yang-Ming University, Taipei City, Taiwan

[5]Health Data Science Research Group, National Taiwan University Hospital, Taipei, Taiwan

[6]The Centre for Intelligent Healthcare, National Taiwan University Hospital, Taipei, Taiwan

[7]Department of Emergency Medicine, National Taiwan University Hospital, Taipei, Taiwan

Corresponding Author:
Chien-Chang Lee, SCD, MD
Department of Emergency Medicine, National Taiwan University Hospital
No 7, Chung-Shan South Road
Taipei, 100
Taiwan
Phone: 886 2 23123456 ext 63485
Fax: 886 2 23223150
Email: hit3transparency@gmail.com

Abstract

Background: Diabetic retinopathy (DR), whose standard diagnosis is performed by human experts, has high prevalence and requires a more efficient screening method. Although machine learning (ML)–based automated DR diagnosis has gained attention due to recent approval of IDx-DR, performance of this tool has not been examined systematically, and the best ML technique for use in a real-world setting has not been discussed.

Objective: The aim of this study was to systematically examine the overall diagnostic accuracy of ML in diagnosing DR of different categories based on color fundus photographs and to determine the state-of-the-art ML approach.

Methods: Published studies in PubMed and EMBASE were searched from inception to June 2020. Studies were screened for relevant outcomes, publication types, and data sufficiency, and a total of 60 out of 2128 (2.82%) studies were retrieved after study selection. Extraction of data was performed by 2 authors according to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), and the quality assessment was performed according to the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2). Meta-analysis of diagnostic accuracy was pooled using a bivariate random effects model. The main outcomes included diagnostic accuracy, sensitivity, and specificity of ML in diagnosing DR based on color fundus photographs, as well as the performances of different major types of ML algorithms.

Results: The primary meta-analysis included 60 color fundus photograph studies (445,175 interpretations). Overall, ML demonstrated high accuracy in diagnosing DR of various categories, with a pooled area under the receiver operating characteristic (AUROC) ranging from 0.97 (95% CI 0.96-0.99) to 0.99 (95% CI 0.98-1.00). The performance of ML in detecting more-than-mild DR was robust (sensitivity 0.95; AUROC 0.97), and by subgroup analyses, we observed that robust performance of ML was not limited to benchmark data sets (sensitivity 0.92; AUROC 0.96) but could be generalized to images collected in clinical practice (sensitivity 0.97; AUROC 0.97). Neural network was the most widely used method, and the subgroup analysis revealed a pooled AUROC of 0.98 (95% CI 0.96-0.99) for studies that used neural networks to diagnose more-than-mild DR.

Conclusions: This meta-analysis demonstrated high diagnostic accuracy of ML algorithms in detecting DR on color fundus photographs, suggesting that state-of-the-art, ML-based DR screening algorithms are likely ready for clinical applications. However, a significant portion of the earlier published studies had methodology flaws, such as the lack of external validation and presence of spectrum bias. The results of these studies should be interpreted with caution.

(J Med Internet Res 2021;23(7):e23863) doi: 10.2196/23863

KEYWORDS

machine learning; diabetic retinopathy; diabetes; deep learning; neural network; diagnostic accuracy

Introduction

Diabetic retinopathy (DR) is the leading cause of vision impairment and blindness among working-aged people in the world [1]. Approximately one-third of people with diabetes mellitus have signs of DR, among whom one-third have vision-threatening DR (VTDR). A meta-analysis estimated global prevalence of any DR and proliferative diabetic retinopathy (PDR) among patients with diabetes to be 35.4% and 7.5%, respectively [2]. The number of patients with DR is approximately 93 million and is expected to rise to 191 million by 2030, as type 2 diabetes has attained the status of a global pandemic, spreading from affluent industrialized nations to the developing world [3].

Vision impairment due to DR can be significantly reduced if diagnosed in early stages and treated appropriately [4]. However, fewer than 60% of patients with diabetes undergo regular eye examinations at intervals recommended by guidelines due to the high cost and low accessibility of ophthalmologic services [3]. The number of people with diabetes that need regular eye examinations has quadrupled in the past three decades. Therefore, the development of an automatic, low-cost, accurate eye screening tool has become an important public health issue [5]. The gold standard for DR screening is based on clinical examinations by human clinicians or the analysis of color fundus photographs via telemedicine [6]. However, both approaches are time-consuming, labor-intensive, and prone to inconsistency due to inherent human subjectivity [7]. Automated systems that are capable of interpreting color fundus photographs with high sensitivity and specificity are critical for widespread implementation of DR screening, and the rise of artificial intelligence (AI), specifically machine learning (ML), has made such automated approaches a possibility.

ML uses existing data to train a computer to recognize a specific pattern or predict a specific outcome in a new data set [6]. Exploration of automated image analysis can be dated back to 1980, when classical ML methods, such as support vector machines and random forests, were used to detect predefined features [8]. These early ML techniques for detecting DR employed mathematical image transformation techniques and image engineering guided by expert-designed rules [9]. The accuracy of this type of analysis did not reach the standard of clinical application. In recent years, the advent of deep learning (DL), a subtype of ML, has transformed the field of automated image analysis [10]. Briefly, DL methods are representation learning methods that use multilayered neural networks, the performance of which can be enhanced by reiteratively changing the internal parameters [11,12]. Unlike other ML approaches, DL does not require image engineering. It develops its own representations needed for pattern recognition after being fed raw data and has shown superior accuracy as compared with other classical ML algorithms [13,14].

Although ML has garnered significant attention with the recent US Food and Drug Administration (FDA) approval of the first ML-based, fully automatic DR screening machine in April 2018 [15], skepticism within the medical community remains regarding the robustness of ML techniques in real-world clinical applications. Given that ophthalmology is among the medical disciplines that have reaped the most benefits from recent AI advancements and that DR screening is one of the most promising ML applications in ophthalmology, we have set out to systematically survey, through meta-analysis, the current status of ML as applied in DR screening based on color fundus photographs. Specifically, we have examined the range of performances reported by different studies and have determined which ML technique is superior for this clinical purpose.

Methods

Search Methods for Identifying Studies

This meta-analysis was performed in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [16]. A literature search for relevant studies published through June 2020 was performed with 2 publicly available databases, PubMed and EMBASE. There were 3 stages to the literature search. No language or population filters were applied, while nonhuman experiments, case reports, guidelines, conference papers, letters, editorials, and review articles were excluded. Filter for publication year was applied only in the second and third stages of the literature search in order to avoid overlapping of search results. Duplicated references in different stages of the literature search were manually excluded. The major search key combination terms were “diabetic retinopathy” OR “diabetic macular edema” OR “macular edema” OR “retinopathy” OR “neovascularized retinopathy” OR “proliferative retinopathy” OR “referable diabetic retinopathy” OR “diabetic macular oedema” OR “proliferative diabetic retinopathy” OR “retinal disorders” OR “diabetic eye disease” OR “vision loss” OR “retinal diseases” OR “macular disease” OR “macular degeneration” OR “macular disorders” crossed with “artificial intelligence” OR “deep learning” OR “transfer learning” OR “machine learning” OR “deep learning system”. The detailed search strategy is provided in Multimedia Appendix 1.

Eligibility Criteria for Considering Studies for This Review

We included studies that evaluated ML algorithms on the accuracy of automated image analysis for screening or diagnosis of DR. We included studies that detected pathological findings of DR, diagnosed DR status, and staged DR severity.

Study Selection

The study selection and data extraction were independently performed by 2 authors (JHW and CCL). After duplicates were removed, titles and abstracts were screened for exclusion of

studies with potentially nonrelevant outcome or publication types and studies applying information other than images in analytical work. When there were multiples studies derived from the same cohort with overlapping study periods, earlier studies were considered duplicates and only the study with the most recent result was included. Retrieved studies with accessible full articles then underwent full-text review. Discrepancies between the reviewers were resolved first by a consensus meeting and then arbitration by a third reviewer if consensus could not be reached.

Data Collection

Data extraction was performed on studies selected after full-text review. A thorough review of each article was performed with the following variables extracted: first author, published year, country, algorithm, image modality, total image size, relevant image size, number of participants and eyes, number of diseased participants and eyes, databases and characteristics, and the sensitivity and specificity of both training and validation sessions. When multiple algorithms were tested on one data set, only the data of the best-performing algorithm were included. Algorithms applied were further classified into 4 main categories: support vector machine (SVM), neural network (NN), random forest (RF), and others. After data extraction, studies were classified into different outcomes of DR, including any DR, more-than-mild diabetic retinopathy (mtmDR), vision-threatening diabetic retinopathy (VTDR), diabetic macular edema, and proliferative diabetic retinopathy (PDR). Studies with relevant data were examined for sufficiency to construct a 2 × 2 contingency table before quality assessment.

Risk of Bias Assessment

The quality of eligible studies was independently assessed by 2 reviewers using the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) tool, which is composed of 4 domains assessing both risk of bias and applicability of clinical practice: patient selection, index test, reference standard, and flow and timing. For each diagnostic study, we determined the risk for bias and general applicability in all 4 domains of QUADAS-2 and reported them separately. A study was considered to have a low risk of bias in one domain if at least half of the variables extracted from the validation session met the requirements of QUADAS-2.

Data Synthesis and Analysis

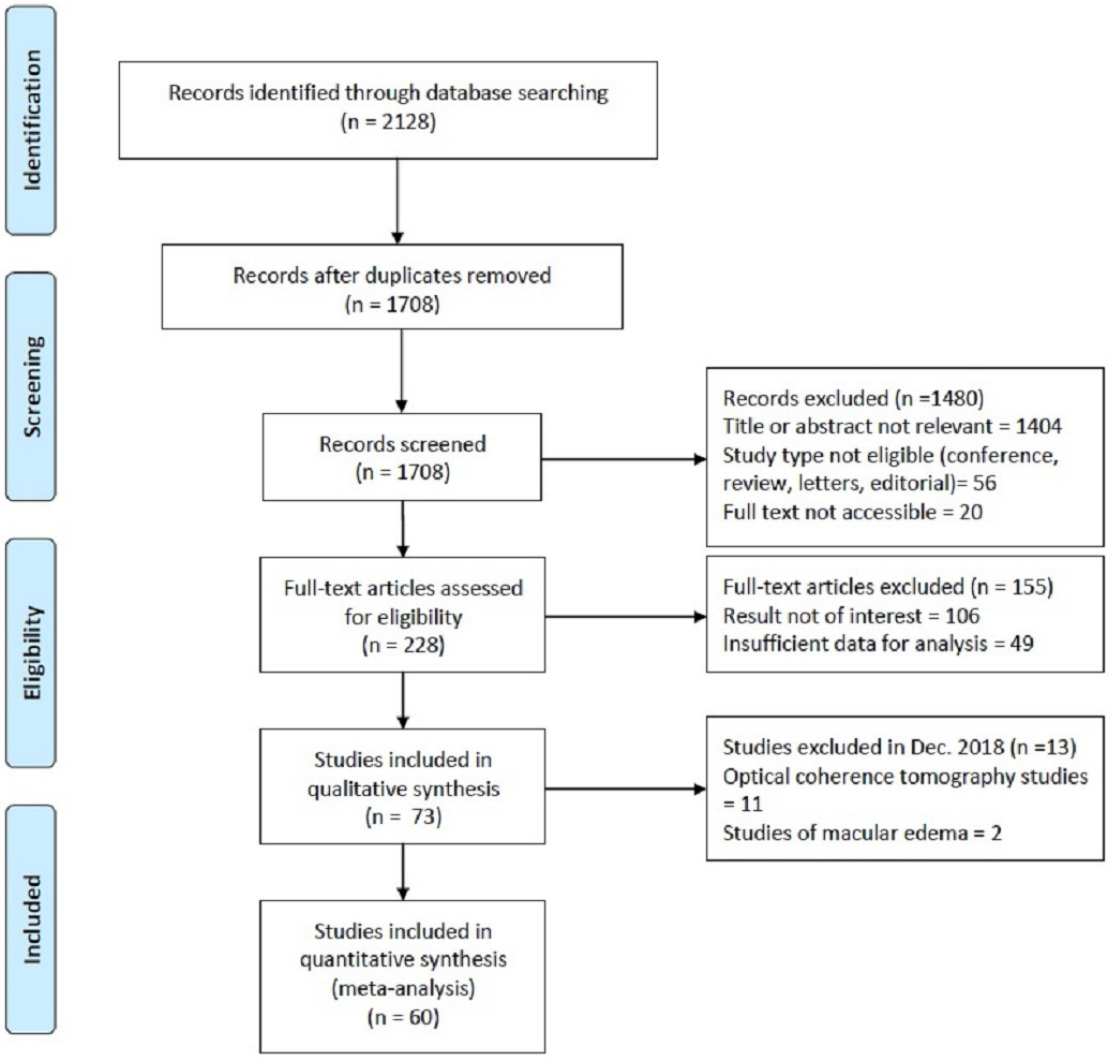
Meta-analysis for diagnostic accuracy of each ML algorithm or DR outcome was performed with a bivariate random effects model to account for both within- and between-study heterogeneity. Results were summarized by using hierarchical summary receiver operating characteristic (ROC) plots and coupled forest plots. Pooled sensitivity, specificity, area under the curve, and positive and negative likelihood ratios were calculated. The bivariate model approach modeled the logit-transformed sensitivity and specificity simultaneously to account for the inherent negative correlation between sensitivity and specificity that might have arisen due to different thresholds in different studies. Heterogeneity was tested using the Cochran Q statistic (P<.05) and quantified with the I^2 statistic, which describes the variation of effect size that is attributable to heterogeneity across studies. For direct clinical interpretation, we also calculated the posttest probability for each type of lesion. We used the prevalence of lesions in the pooled study population as the informative prior, and derived the posttest probability of each type of lesion based on the pooled positive and negative likelihood ratios. Results are presented as Fagan nomograms. The presence and effect of publication bias were examined with Deeks tests. When publication bias is present, Deeks funnel plots are usually asymmetrical. We used a trim-and-fill method to impute hypothetical missing studies due to publication bias. Trim-and-fill odds ratios (ORs) were reported when the tests for publication bias were significant. We performed sensitivity analyses to examine the potential effects of different demographic factors, ML algorithms, and types of training or validating databases. All analyses except for the summary ROC curve were conducted by the “mada” package in R software (The R Foundation for Statistical Computing). Summary ROC and area under the ROC (AUROC) were calculated by the “midas” package in Stata 14.0 (StataCorp). A 2-sided P value <0.05 indicated statistical significance for all tests.

Results

Search Results

The first 2 stages of the literature search (performed in May 2018 and December 2018) yielded 668 hits from PubMed and 809 hits from EMBASE. After screening titles and abstracts, we excluded 336 studies for duplication and 941 studies for nonrelevant abstracts or publication types. A total of 187 studies went through full-text review, 99 of which were excluded for a result that was not of interest and 43 of which were excluded for insufficient data for analysis. After completing qualitative synthesis of the 45 studies, we proceeded to only include ML studies that involved use of color fundus photograph for DR screening. After further exclusion, only 32 studies were retrieved after the second stage of literature selection [13,15,17-46]. The third stage of the literature search was performed in June 2020, and 28 out of 651 studies were retrieved after literature selection [47-74], resulting in a total of 60 (N=32+28) included studies for final analysis. A new category composed of VTDR and PDR (VTDR+PDR) was created for examination of diagnostic accuracy of the most treatment-urgent group. A flowchart of the literature search and study selection process is summarized in Figure 1.

Figure 1. Flowchart of study selection.



Study Characteristics

Multimedia Appendix 2 Table S1 summarizes the study-level characteristics of studies assessing the diagnostic accuracy of ML algorithms for different categories of DR. Of the 60 studies, 35 studies (58%) evaluated any DR, 23 (38%) mtmDR, 12 (20%) VTDR, and 12 (20%) PDR. Publicly available benchmark databases, such as Messidor, Structured Analysis of the Retina (STARE), Digital Retinal Images for Vessel Extraction (DRIVE), DIARETDB, e-Ophtha, and EyePACs were used for testing of the ML algorithms in 40 of the 60 (67%) studies. The characteristics of these publicly available retinal image databases are summarized in Multimedia Appendix 3. The distribution of categories of ML algorithms used was as follows: SVM (6/60, 10%), RF (2/60, 3%), NN (37/60, 62%), and others (17/60, 28%). The general principles of these ML algorithms are described in Multimedia Appendix 4.

Quality Assessment

Quality assessments using the QUADAS-2 criteria are summarized in Multimedia Appendix 5. Most studies (56/60, 93%) presented a clear source of patient recruitment or selection criteria and processes, and were at a low risk for bias. Of the 60 studies, 3 (5%) reported limited information on the establishment of reference standard and were at a high risk for bias, and 4 (7%) reported insufficient blinding to a reference standard during interpretation of the index test results and were at a high risk for bias. For study applicability, 1 study (2%) in the index test section and 4 (7%) studies in the patient selection section were recorded to be at a high risk of concern, due to insufficient information reported.

Synthesis of Results

A summary of data of included studies is presented in Multimedia Appendix 6. Pooled sensitivities, specificities, likelihood ratios, AUROCs, and I^2 statistics for the 5 DR categories, including any DR, mtmDR, VTDR, PDR, and VTDR+PDR, are presented in Table 1. As some studies might have used more than 2 data sets for validation, performance of ML derived from each data set was viewed as individual data, and we used “data” as the unit for calculation (eg, 35 included studies performed evaluation of ML on identifying any DR, resulting in a total of 53 data for synthesis and analysis). The

hierarchical summary ROC plots for the 4 main DR categories, including DR, mtmDR, VTDR, and PDR, are also presented (Figures 2-5). ML showed a high overall accuracy in detecting the 5 categories of DR, with a pooled AUROC ranging from 0.97 (95% CI 0.96-0.99) for mtmDR and VTDR+PDR to 0.99 (95% CI 0.98-1.00) for VTDR and PDR. The pooled sensitivity for all 5 categories was high, ranging from 0.93 (95% CI 0.87-0.96) for PDR to 0.97 (95% CI 0.94-0.99) for VTDR. The pooled specificity, however, showed more variation: from 0.90 (95% CI 0.87-0.93) for mtmDR to 0.98 (95% CI 0.96-0.99) for PDR. The Fagan plots for different DR categories are presented in Multimedia Appendix 7. For images that were classified as positive by the ML algorithms, the posttest probability for DR, mtmDR, VTDR, and PDR was 87%, 71%, 66%, and 77%, respectively. For images that were classified as negative by the ML algorithms, the posttest probability for DR, mtmDR, VTDR, and PDR was 4%, 1%, 0%, and 1%, respectively.

Table 1. Pooled analysis for diagnostic accuracy of diabetic retinopathy by machine learning on color fundus photographs.

Goal of detection	Data[a], n	Sen[b,c]	Spe[d,c]	LR+[e,c]	LR–[f,c]	AUROC[g,c]	I^2 statistic[c]	Publication bias (*P* value)
Any DR[h]	53	0.94 (0.91-0.96)	0.92 (0.88-0.95)	12.4 (8.0-19.3)	0.07 (0.05-0.09)	0.98 (0.96-0.99)	32 (22-42)	.01
mtmDR[i]	40	0.95 (0.93-0.97)	0.90 (0.87-0.93)	9.7 (7.4-12.7)	0.05 (0.04-0.08)	0.97 (0.96-0.99)	29 (18-40)	.11
VTDR[j]	15	0.97 (0.94-0.99)	0.94 (0.87-0.98)	17.3 (7.5-39.9)	0.03 (0.01-0.06)	0.99 (0.98-1.00)	32 (9-56)	.33
PDR[k]	22	0.93 (0.87-0.96)	0.98 (0.96-0.99)	38.5 (21.7-68.4)	0.07 (0.04-0.13)	0.99 (0.98-1.00)	29 (11-46)	.11
VTDR and PDR	37	0.96 (0.93-0.98)	0.97 (0.94-0.98)	24.3 (14.5-38.5)	0.07 (0.05-0.10)	0.97 (0.96-0.99)	N/A	.06

[a]Machine learning data derived from each data set was viewed as individual data, and we used “data” as the unit for calculation.

[b]Sen: sensitivity.

[c]Values in this column are as follows: mean (95% confidence interval).

[d]Spe: specificity.

[e]LR+: positive likelihood ratio.

[f]LR–: negative likelihood ratio.

[g]AUROC: area under the receiver operating characteristic.

[h]DR: diabetic retinopathy.

[i]mtmDR: more-than-mild diabetic retinopathy.

[j]VTDR: vision-threatening diabetic retinopathy.

[k]PDR: proliferative diabetic retinopathy.

Figure 2. SROC curves for diagnosis of any diabetic retinopathy on color fundus photographs. AUC: area under the curve; Sens: sensitivity; Spec: specificity; SROC: summary receiver operating characteristics.

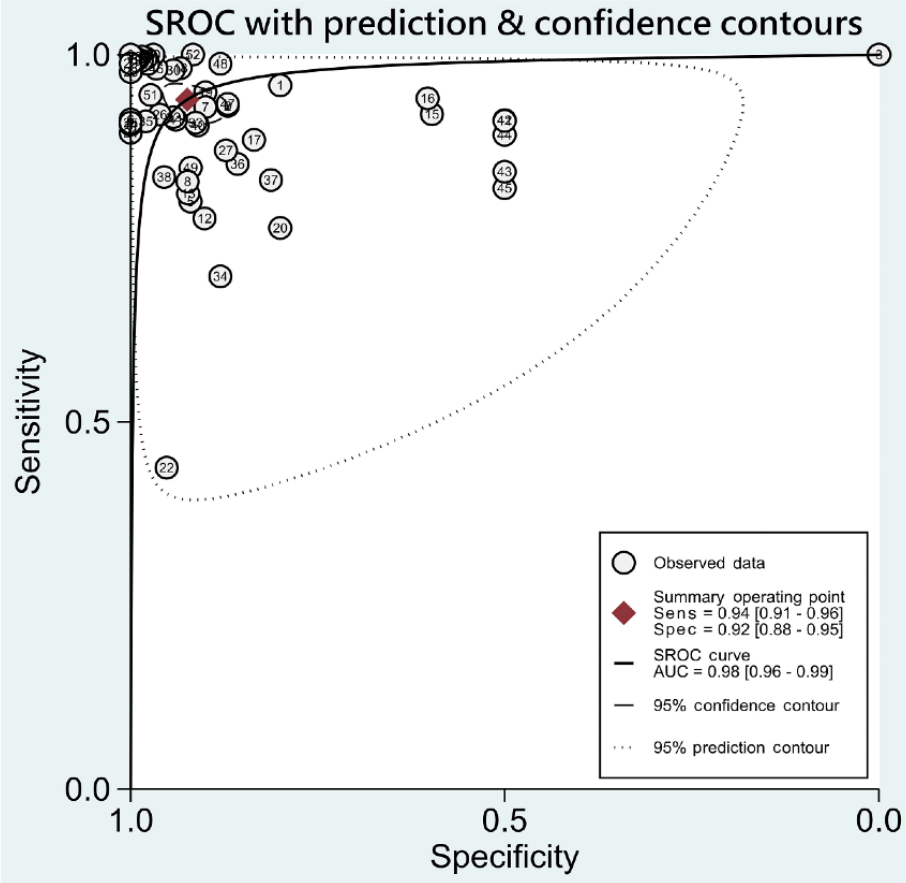


Figure 3. SROC curves for diagnosis of more-than-mild diabetic retinopathy on color fundus photographs. Sens: sensitivity; Spec: specificity; SROC: summary receiver operating characteristics; AUC: area under the curve.

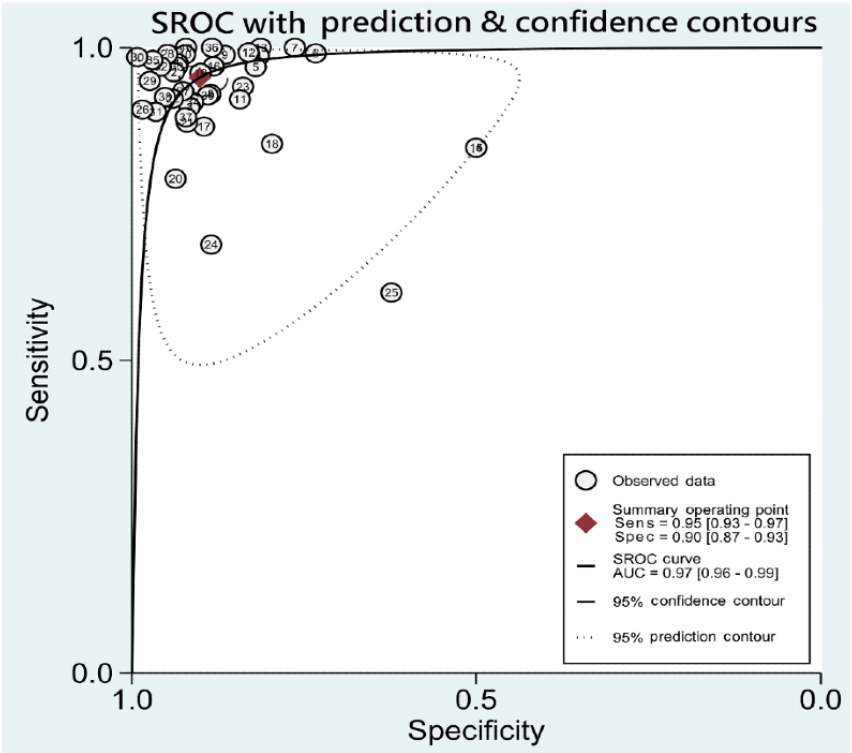


Figure 4. SROC curves for diagnosis of vision-threatening diabetic retinopathy on color fundus photographs. AUC: area under the curve; Sens: sensitivity; Spec: specificity; SROC: summary receiver operating characteristics.

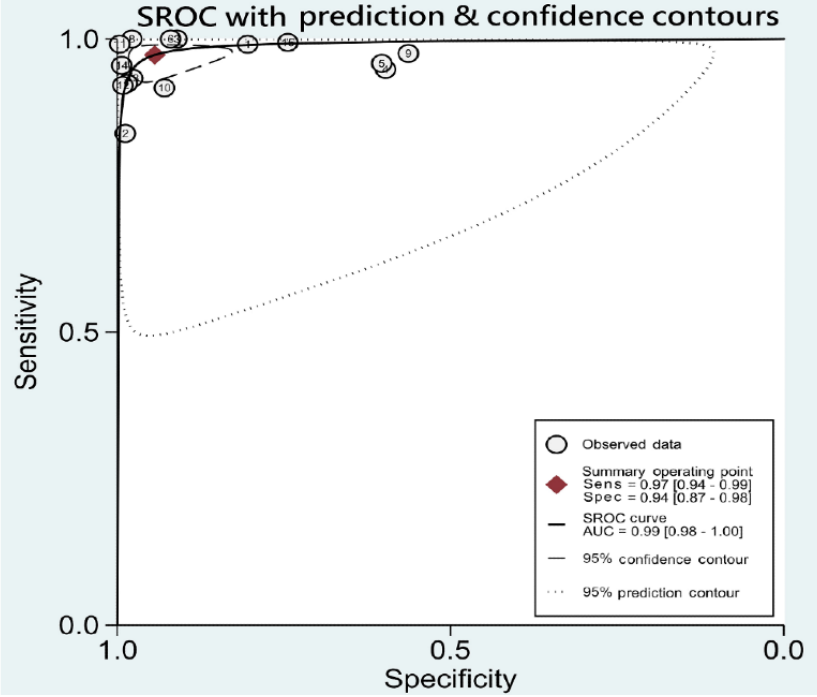
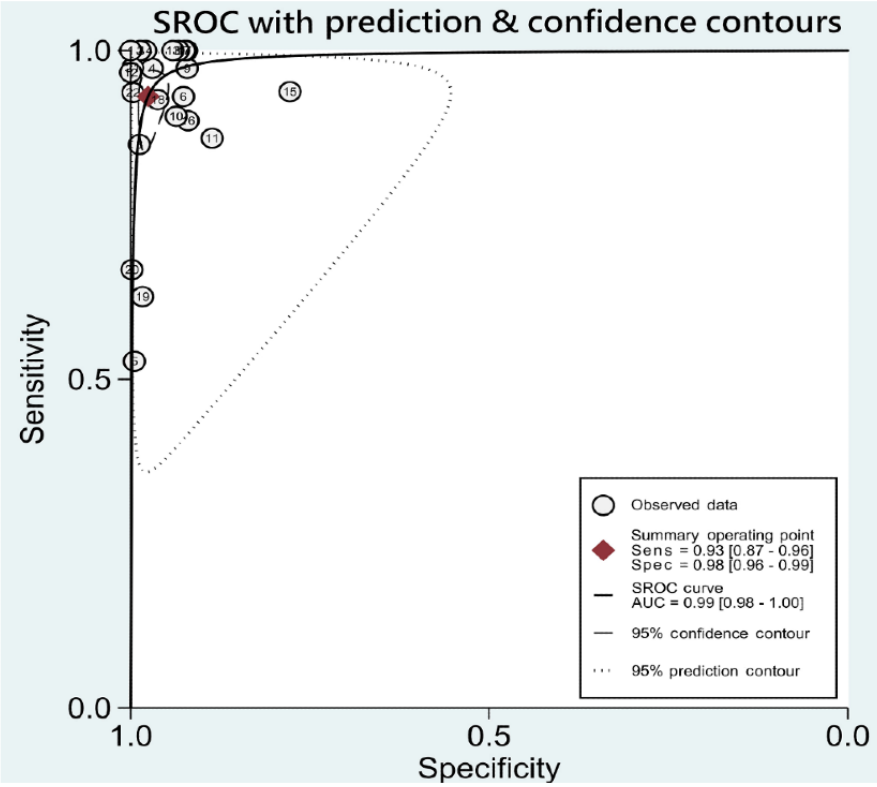


Figure 5. SROC curves for diagnosis of proliferative diabetic retinopathy on color fundus photographs. AUC: area under the curve; Sens: sensitivity; Spec: specificity; SROC: summary receiver operating characteristics.



Subgroup Analyses and Sensitivity Analysis

We performed subgroup analyses for mtmDR studies to explore the possible heterogeneity in test accuracy (Table 2). The main causes of heterogeneity included in the analysis were algorithm type, mean age of subject populations, and validation set selection. For this subgroup analysis, 23 studies were included, with a total of 40 data obtained from different testing data sets. Of the 23 studies, the 22 studies that applied NN algorithms demonstrated high pooled performance (summary AUROC 0.98; 95% CI 0.96-0.99), sensitivity (sensitivity 0.95; 95% CI 0.93-0.97), and specificity (specificity 0.91; 95% CI 0.88-0.93). The only study that used a different kind of ML algorithm (instance learning) reported significantly inferior sensitivity

(0.84) under preset specificity (0.50). Of the 60 studies, 19 (83%) tested the algorithm's performance on data sets with subject populations with a mean age greater than 50 years. Pooled sensitivity of data from these studies was high (sensitivity 0.95; 95% CI 0.93-0.97), and the pooled specificity was moderate (specificity 0.89; 95% CI 0.85-0.92). Compared with algorithms that used benchmark data sets for validation (pooled sensitivity 0.92; 95% CI 0.87-0.95), the pooled sensitivities of algorithms validated by clinical data sets (sensitivity 0.97; 95% CI 0.95-0.98) and independent data sets (sensitivity 0.96; 95% CI 0.93-0.97) were not inferior. The results of pooled AUROCs validated by these 3 types of data sets were similar, implying that robust performance of ML algorithms can be generalized to images collected in clinical practice.

Table 2. Subgroup analysis for diagnostic accuracy of mtmDR retinopathy on color fundus photographs.

Features of subgroup	Data[a], n	Sen[b,c]	Spe[d,c]	LR+[e,c]	LR–[f,c]	AUROC[g,c]	I^2 statistic[c]	Publication bias (P value)
Overall mtmDR[h]	40	0.95 (0.93, 0.97)	0.90 (0.87, 0.93)	9.7 (7.4, 12.7)	0.05 (0.04, 0.08)	0.97 (0.96, 0.99)	29 (18, 40)	.11
Mean age> 50 years	32	0.95 (0.93, 0.97)	0.89 (0.85, 0.92)	8.8 (6.4, 12.0)	0.05 (0.03, 0.08)	0.97 (0.95, 0.98)	33 (20, 46)	.22
NN[i] algorithms	38	0.95 (0.93, 0.97)	0.91 (0.88, 0.93)	10.1 (7.7, 13.2)	0.05 (0.03, 0.07)	0.98 (0.96, 0.99)	30 (19, 41)	.14
Benchmark test sets	15	0.92 (0.87, 0.95)	0.90 (0.82, 0.94)	9.0 (4.8, 16.6)	0.09 (0.05, 0.16)	0.96 (0.94, 0.98)	25 (10, 39)	.22
Clinical Test sets	25	0.97 (0.95, 0.98)	0.90 (0.88, 0.92)	10.0 (7.9, 12.6)	0.04 (0.02, 0.06)	0.97 (0.96, 0.98)	30 (15, 45)	.06
External validation	31	0.96 (0.93, 0.97)	0.90 (0.87, 0.93)	9.7 (7.2, 13.0)	0.05 (0.03, 0.07)	0.98 (0.96, 0.99)	29 (16, 42)	.08

[a]Machine learning data derived from each data set was viewed as individual data, and we used "data" as the unit for calculation.

[b]Sen: sensitivity.

[c]Values in this column are as follows: mean (95% confidence interval).

[d]Spe: specificity.

[e]LR+: positive likelihood ratio.

[f]LR–: negative likelihood ratio.

[g]AUROC: area under the receiver operating characteristic.

[h]mtmDR: more-than-mild diabetic retinopathy.

[i]NN: neural network.

Publication Bias

The test for publication bias was generally not significant in different categories of DR (Deeks test P=.01; Multimedia Appendix 8), except for any DR. Trim-and-fill analysis showed the diagnostic OR remained insignificant (OR 0.50, 95% CI 0.25-1.01) after hypothetical unpublished data were included for analysis (Multimedia Appendix 9).

Discussion

Principal Results and Comparison With Prior Work

This systematic review synthesizes the available evidence and compares the diagnostic accuracy of ML algorithms for the detection of DR based on color fundus photographs. The primary meta-analysis included 60 studies with 445,175 interpretations. Out of the 60 studies, 35 (58%) were validated by external testing data sets that were completely independent of the training data sets. Overall, ML demonstrated a robust performance in detecting different DR categories, with a pooled sensitivity of 0.93 to 0.97 and a pooled specificity of 0.90 to 0.98. The pooled sensitivity compares favorably to reported sensitivities of 73% [75], 34% [76], and 33% [77] achieved by board-certified ophthalmologists performing indirect ophthalmoscopy and to reported sensitivities of 92% [78] and 89% [79] achieved by ophthalmologists interpreting digital fundus photographs. Our analysis suggests that the performance of ML algorithms in detecting DR based on color fundus photographs is likely to be on par with human clinicians and supports a previous study that compared humans head to head with ML. Rajalakshmi et al [37] compared the performance of an AI DR screening software (EyeArt) on smartphone-based fundus photographs of 296 patients to the performance of human graders who evaluated the same data set. The EyeArt achieved a high sensitivity of 95.8% for any retinopathy and 99.1% for VTDR, both of which were on par with human graders. Our pooled data suggest that ML techniques are more sensitive than specific in DR detection. It is unclear whether this is a reflection of the limitations of ML techniques for this clinical purpose or whether it is by design. It is possible that model developers of these studies chose optimal statistical thresholds that favored sensitivity over specificity. Regardless, the lower specificity should not pose a major issue, as false negatives are much more problematic than are false positives in the context of disease screening.

Furthermore, the major causes of false positives in retinal image interpretation, including inadequate image quality and artifacts [55], are modifiable with future improvement in image quality control.

To further facilitate direct clinical interpretations, we used Fagan nomograms to determine whether a patient with a positive or negative finding by ML actually has that particular finding as per the gold standard. For any DR, in a population with a DR prevalence of 36%, a positive likelihood ratio of 12.4 translates into a posttest probability of 87%. In other words, approximately 9 out of 10 patients with a positive ML diagnosis of DR can be expected to have DR as per the gold standard. The diagnostic value for ML to rule out DR performed as well as its rule-in value. In the same population, a negative ML diagnosis translates into a 4% posttest probability of any DR (negative likelihood ratio 0.07) and only a 1% posttest probability of mtmDR (negative likelihood ratio 0.05). These numbers again suggest that ML is extremely sensitive in detecting overall DR and mtmDR based on color fundus photographs and that the rate of false negatives are low.

We performed an in-depth analysis of studies that involved the detection of mtmDR, as Abramoff et al's [15] pivotal trial involved the detection of mtmDR and led to the FDA's approval of the first fully automated ML system. Among the 16 mtmDR studies conducted by other research teams that were also externally validated, 14 showed performance superior to the preset end points (sensitivity >85%; specificity > 82.5%) used in Abramoff et al's trial. Although only 5 out of these studies were prospectively evaluated in a real-world setting as the Abramoff algorithm was, this suggests that Abramoff et al's trial was no accident and that ML algorithms in general are likely capable of producing clinical grade detections of mtmDR based on color fundus photographs. In addition, no statistically significant difference in pooled AUROC between studies validated by benchmark databases and studies validated by clinical databases was identified within this group.

To the best of our knowledge, previous meta-reviews on DR screening have focused on the performance of DL algorithms alone [80,81]. DL is only a subtype of ML, and other ML techniques, such as SVM and RF, can be used to detect DR as well. Therefore, our meta-analysis was more comprehensive than these previous studies, as it included all ML studies, including DL studies, published through 2020. In addition, the review by Nielsen et al [80] did not conduct pooled analysis on the results of past studies, while our study did. The meta-analysis by Islam et al [81] focused mainly on detection of referable DR, while our study was broader and more fine grained, as it evaluated the ability of ML to detect different categories of DR, including any DR, mtmDR (referrable DR), VTDR, and PDR. These analyses are clinically meaningful, as different categories of DR require different management strategies. For example, while patients with moderate nonproliferative DR (a subset within referrable DR) should be further evaluated by ophthalmologists at some point, patients with VTDR require immediate referrals to retinal specialists.

Use and Predominance of NN Algorithms

NN algorithms, especially deep convolutional NN algorithms, were generally recognized as the best ML technique for automated medical image analysis. NN algorithms were also the most-used technique in diagnosing DR of all categories in our study, being used in 37 of the 60 (62%) studies. As for the 23 studies evaluating mtmDR (Table 2), NN algorithms were used in 22 studies and contributed to the high pooled AUROC of 0.98 (95% CI 0.96-0.99). In addition, we ranked the performance of the included studies by sensitivity, specificity, and quality. The top-5 performing, high-quality (based on QUADUS-2 and study design) studies are listed in Multimedia Appendix 10, and 4 out of the 5 studies used NN algorithms. This result confirms that NN is the cutting-edge ML technique for medical image classification, at least in the context of DR detection.

Limitations

Our study was based on a rigorous literature search, and a validated appraisal tool was used to determine the risk of bias of included studies. Several limitations should be considered, however. First, of the 60 studies included for final analysis, only 35 applied true external validation. For those studies without external validation, the generalizability of their ML algorithms was not adequately evaluated, and thus their reported performance should be interpreted with caution. Second, without sufficient details, we were unable to conduct subgroup analysis on populations with available key factors of DR that could influence the clinical practicability of the diagnostic tool. Bias could have been introduced by poor reporting of patient characteristics of the included studies. Finally, except for Abramoff et al's trial [15] and 5 other prospectively conducted studies [48,52,55,60,66], all other studies on ML-based DR diagnosis were validated by retrospective data. Due to spectrum bias, an overestimation of ML's performance in a real-world setting is possible and should be considered.

Conclusions

ML algorithms for diagnosing DR based on color fundus photographs have shown high diagnostic accuracy for different categories of DR. Specifically, the performances of ML algorithms in detecting mtmDR, the widely accepted threshold for clinically relevant DR, compare favorably to those of clinical examinations by ophthalmologists and to those of expert grading of digital fundus photographs. To the best of our knowledge, this is the first meta-analysis in the published literature that quantitatively assessed the performance of ML algorithms for a specific medical image classification task. As evidence-based medicine expands from therapy to diagnosis, the information from this systematic review may provide important evidence in the determination of the proper and efficacious use of ML algorithms in the diagnosis or screening of DR and may serve as a framework for similar analyses of other medical conditions conducted in the future. However, our meta-analysis also showed that a significant portion of the published studies had methodological flaws, such as the lack of external validation and presence of spectrum bias. Therefore, more rigorous prospective studies would be helpful in establishing the true efficacy of these algorithms in real-life clinical care.

Acknowledgments

This study was funded by the Taiwan National Ministry of Science and Technology (grants MOST 104-2314-B-002 -039 -MY3 and MOST 106-2811-B-002-048). The sponsor or funding organization had no role in the design or conduct of this research.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Search detail for PubMed and EMBASE.
[DOCX File , 33 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Summary of characteristics of the included studies.
[DOCX File , 74 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Characteristics of public databases for benchmarking diabetic retinopathy detection from color fundus photographs.
[DOCX File , 40 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Principles of major machine learning algorithms and methods applied in automated image analysis.
[DOCX File , 39 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Summary of QUADAS-2 assessment for included studies.
[PDF File (Adobe PDF File), 145 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Summary of data of included studies.
[DOCX File , 81 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Fagan plot for diagnosis of different categories of diabetic retinopathy on color fundus photographs.
[PDF File (Adobe PDF File), 728 KB-Multimedia Appendix 7]

Multimedia Appendix 8

Deeks funnel plot for 4 main types of diabetic retinopathy lesions on color fundus photograph.
[PDF File (Adobe PDF File), 374 KB-Multimedia Appendix 8]

Multimedia Appendix 9

A filled funnel plot that filled the potential missing studies (square) due to publication bias.
[DOCX File , 264 KB-Multimedia Appendix 9]

Multimedia Appendix 10

Characteristics and best results of 5 high-quality studies with top performances.
[DOCX File , 46 KB-Multimedia Appendix 10]

References

1. Fong DS, Aiello L, Gardner TW, King GL, Blankenship G, Cavallerano JD, American Diabetes Association. Diabetic retinopathy. Diabetes Care 2003 Jan;26(1):226-229. [doi: 10.2337/diacare.26.1.226] [Medline: 12502685]

2. Yau JWY, Rogers SL, Kawasaki R, Lamoureux EL, Kowalski JW, Bek T, et al. Global prevalence and major risk factors of diabetic retinopathy. Diabetes Care 2012 Mar;35(3):556-564 [FREE Full text] [doi: 10.2337/dc11-1909] [Medline: 22301125]
3. Zheng Y, He M, Congdon N. The worldwide epidemic of diabetic retinopathy. Indian J Ophthalmol 2012;60(5):428-431. [doi: 10.4103/0301-4738.100542] [Medline: 22944754]
4. Vashist P, Singh S, Gupta N, Saxena R. Role of early screening for diabetic retinopathy in patients with diabetes mellitus: an overview. Indian J Community Med 2011 Oct;36(4):247-252 [FREE Full text] [doi: 10.4103/0970-0218.91324] [Medline: 22279252]
5. Zheng Y, Ley SH, Hu FB. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. Nat Rev Endocrinol 2018 Feb;14(2):88-98. [doi: 10.1038/nrendo.2017.151] [Medline: 29219149]
6. Viswanath K, McGavin DDM. Diabetic retinopathy: clinical findings and management. Community Eye Health 2003;16(46):21-24 [FREE Full text] [Medline: 17491851]
7. Heitmar R, Kalitzeos AA, Patel SR, Prabhu-Das D, Cubbidge RP. Comparison of subjective and objective methods to determine the retinal arterio-venous ratio using fundus photography. J Optom 2015;8(4):252-257 [FREE Full text] [doi: 10.1016/j.optom.2014.07.002] [Medline: 26386537]
8. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018 Dec;18(8):500-510 [FREE Full text] [doi: 10.1038/s41568-018-0016-5] [Medline: 29777175]
9. Wernick MN, Yang Y, Brankov JG, Yourganov G, Strother SC. Machine learning in medical imaging. IEEE Signal Process Mag 2010 Jul;27(4):25-38 [FREE Full text] [doi: 10.1109/MSP.2010.936730] [Medline: 25382956]
10. Wong TY, Bressler NM. Artificial intelligence with deep learning technology looks into diabetic retinopathy screening. JAMA 2016 Dec 13;316(22):2366-2367 [FREE Full text] [doi: 10.1001/jama.2016.17563] [Medline: 27898977]
11. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 28;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]
12. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017 Dec;2(4):230-243 [FREE Full text] [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
13. 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 Dec 13;316(22):2402-2410. [doi: 10.1001/jama.2016.17216] [Medline: 27898976]
14. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019 Jan;25(1):24-29. [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
15. 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 [FREE Full text] [doi: 10.1038/s41746-018-0040-6] [Medline: 31304320]
16. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev 2015 Jan;4:1 [FREE Full text] [doi: 10.1186/2046-4053-4-1] [Medline: 25554246]
17. 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 Nov;55(11):1959-1974. [doi: 10.1007/s11517-017-1638-6] [Medline: 28353133]
18. 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 Oct 01;57(13):5200-5206. [doi: 10.1167/iovs.16-19964] [Medline: 27701631]
19. Adal KM, Sidibé D, Ali S, Chaum E, Karnowski TP, Mériaudeau F. Automated detection of microaneurysms using scale-adapted blob analysis and semi-supervised learning. Comput Methods Programs Biomed 2014 Apr;114(1):1-10. [doi: 10.1016/j.cmpb.2013.12.009] [Medline: 24529636]
20. Agurto C, Barriga ES, Murray V, Nemeth S, Crammer R, Bauman W, et al. Automatic detection of diabetic retinopathy and age-related macular degeneration in digital fundus images. Invest Ophthalmol Vis Sci 2011 Jul 29;52(8):5862-5871. [doi: 10.1167/iovs.10-7075] [Medline: 21666234]
21. Annie GVG, Kaja MS. Diabetic retinopathy screening system: A validation analysis with multiple fundus image databases. Biomedical Research (India) 2017;28(4):A.
22. Chaum E, Karnowski TP, Govindasamy VP, Abdelrahman M, Tobin KW. Automated diagnosis of retinopathy by content-based image retrieval. Retina 2008;28(10):1463-1477. [doi: 10.1097/IAE.0b013e31818356dd] [Medline: 18997609]
23. Ganesan K, Martis RJ, Acharya UR, Chua CK, Min LC, Ng EYK, et al. Computer-aided diabetic retinopathy detection using trace transforms on digital fundus images. Med Biol Eng Comput 2014 Aug;52(8):663-672. [doi: 10.1007/s11517-014-1167-5] [Medline: 24958614]
24. Gardner GG, Keating D, Williamson TH, Elliott AT. Automatic detection of diabetic retinopathy using an artificial neural network: a screening tool. Br J Ophthalmol 1996 Nov;80(11):940-944 [FREE Full text] [doi: 10.1136/bjo.80.11.940] [Medline: 8976718]

25. Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. Ophthalmology 2017 Dec;124(7):962-969. [doi: 10.1016/j.ophtha.2017.02.008] [Medline: 28359545]
26. Gupta G, Kulasekaran S, Ram K, Joshi N, Sivaprakasam M, Gandhi R. Local characterization of neovascularization and identification of proliferative diabetic retinopathy in retinal fundus images. Comput Med Imaging Graph 2017 Jan;55:124-132. [doi: 10.1016/j.compmedimag.2016.08.005] [Medline: 27634547]
27. Jelinek HF, Cree MJ, Leandro JJG, Soares JVB, Cesar RM, Luckie A. Automated segmentation of retinal blood vessels and identification of proliferative diabetic retinopathy. J Opt Soc Am A Opt Image Sci Vis 2007 May;24(5):1448-1456. [doi: 10.1364/josaa.24.001448] [Medline: 17429492]
28. Malathi K, Nedunchelian R. A recursive support vector machine (RSVM) algorithm to detect and classify diabetic retinopathy in fundus retina images. biomedicalresearch 2018:57-64. [doi: 10.4066/biomedicalresearch.29-16-2328]
29. Fadafen MK, Mehrshad N, Razavi SM. Detection of diabetic retinopathy using computational model of human visual system. biomedicalresearch 2018;29(9):1956-1960. [doi: 10.4066/biomedicalresearch.29-18-551]
30. Li Z, Keel S, Liu C, He Y, Meng W, Scheetz J, et al. An automated grading system for detection of vision-threatening referable diabetic retinopathy on the basis of color fundus photographs. Diabetes Care 2018 Dec;41(12):2509-2516. [doi: 10.2337/dc18-0147] [Medline: 30275284]
31. Bala MP, Vijayachitra S. Extraction of retinal blood vessels and diagnosis of proliferative diabetic retinopathy using extreme learning machine. J Med Imaging Hlth Inform 2015 Apr 01;5(2):248-256. [doi: 10.1166/jmihi.2015.1380]
32. Orlando JI, Prokofyeva E, Del Fresno M, Blaschko MB. An ensemble deep learning based approach for red lesion detection in fundus images. Comput Methods Programs Biomed 2018 Jan;153:115-127. [doi: 10.1016/j.cmpb.2017.10.017] [Medline: 29157445]
33. Orlando JI, van Keer K, Barbosa Breda J, Manterola HL, Blaschko MB, Clausse A. Proliferative diabetic retinopathy characterization based on fractal features: Evaluation on a publicly available dataset. Med Phys 2017 Dec;44(12):6425-6434. [doi: 10.1002/mp.12627] [Medline: 29044550]
34. Pires R, Carvalho T, Spurling G, Goldenstein S, Wainer J, Luckie A, et al. Automated multi-lesion detection for referable diabetic retinopathy in indigenous health care. PLoS One 2015;10(6):e0127664 [FREE Full text] [doi: 10.1371/journal.pone.0127664] [Medline: 26035836]
35. Quellec G, Charrière K, Boudi Y, Cochener B, Lamard M. Deep image mining for diabetic retinopathy screening. Med Image Anal 2017 Jul;39:178-193 [FREE Full text] [doi: 10.1016/j.media.2017.04.012] [Medline: 28511066]
36. Quellec G, Lamard M, Abràmoff MD, Decencière E, Lay B, Erginay A, et al. A multiple-instance learning framework for diabetic retinopathy screening. Med Image Anal 2012 Aug;16(6):1228-1240. [doi: 10.1016/j.media.2012.06.003] [Medline: 22850462]
37. Rajalakshmi R, Subashini R, Anjana RM, Mohan V. Automated diabetic retinopathy detection in smartphone-based fundus photography using artificial intelligence. Eye (Lond) 2018 Jun;32(6):1138-1144 [FREE Full text] [doi: 10.1038/s41433-018-0064-9] [Medline: 29520050]
38. Raju M, Pagidimarri V, Barreto R, Kadam A, Kasivajjala V, Aswath A. Development of a deep learning algorithm for automatic diagnosis of diabetic retinopathy. Stud Health Technol Inform 2017;245:559-563. [Medline: 29295157]
39. Ramachandran N, Hong SC, Sime MJ, Wilson GA. Diabetic retinopathy screening using deep neural network. Clin Exp Ophthalmol 2018 May;46(4):412-416. [doi: 10.1111/ceo.13056] [Medline: 28881490]
40. Sangeethaa SN, Uma Maheswari P. An intelligent model for blood vessel segmentation in diagnosing DR using CNN. J Med Syst 2018 Aug 15;42(10):175. [doi: 10.1007/s10916-018-1030-6] [Medline: 30109508]
41. B. Sumathy, S. Poornachandra. Automated DR and prediction of various related diseases of retinal fundus images. biomedicalresearch 2018:325-332. [doi: 10.4066/biomedicalresearch.29-17-480]
42. Ting DSW, Cheung CY, 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 Dec 12;318(22):2211-2223 [FREE Full text] [doi: 10.1001/jama.2017.18152] [Medline: 29234807]
43. Usman Akram M, Khalid S, Tariq A, Younus Javed M. Detection of neovascularization in retinal images using multivariate m-Mediods based classifier. Comput Med Imaging Graph 2013;37(5-6):346-357 [FREE Full text] [doi: 10.1016/j.compmedimag.2013.06.008] [Medline: 23916066]
44. Welikala RA, Fraz MM, Dehmeshki J, Hoppe A, Tah V, Mann S, et al. Genetic algorithm based feature selection combined with dual classification for the automated detection of proliferative diabetic retinopathy. Comput Med Imaging Graph 2015 Jul;43:64-77. [doi: 10.1016/j.compmedimag.2015.03.003] [Medline: 25841182]
45. Yu S, Xiao D, Kanagasingam Y. Machine learning based automatic neovascularization detection on optic disc region. IEEE J. Biomed. Health Inform 2018 May;22(3):886-894. [doi: 10.1109/jbhi.2017.2710201]
46. Zhang Y, An M. An active learning classifier for further reducing diabetic retinopathy screening system cost. Comput Math Methods Med 2016;2016:4345936 [FREE Full text] [doi: 10.1155/2016/4345936] [Medline: 27660645]
47. Badgujar R, Deore P. Hybrid nature inspired SMO-GBM classifier for exudate classification on fundus retinal images. IRBM 2019 Mar;40(2):69-77 [FREE Full text] [doi: 10.1016/j.irbm.2019.02.003] [Medline: 32595044]

48. Bellemo V, Lim ZW, Lim G, Nguyen QD, Xie Y, Yip MYT, et al. Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. The Lancet Digital Health 2019 May;1(1):e35-e44. [doi: 10.1016/s2589-7500(19)30004-4]
49. Bhaskaranand M, Ramachandra C, Bhat S, Cuadros J, Nittala MG, Sadda SR, et al. The value of automated diabetic retinopathy screening with the EyeArt system: a study of more than 100,000 consecutive encounters from people with diabetes. Diabetes Technol Ther 2019 Nov;21(11):635-643 [FREE Full text] [doi: 10.1089/dia.2019.0164] [Medline: 31335200]
50. Chowdhury AR, Chatterjee T, Banerjee S. A Random Forest classifier-based approach in the detection of abnormalities in the retina. Med Biol Eng Comput 2019 Jan;57(1):193-203. [doi: 10.1007/s11517-018-1878-0] [Medline: 30076537]
51. Colomer A, Igual J, Naranjo V. Detection of early signs of diabetic retinopathy based on textural and morphological information in fundus images. Sensors (Basel) 2020 Feb 13;20(4):1005 [FREE Full text] [doi: 10.3390/s20041005] [Medline: 32069912]
52. Gulshan V, Rajan RP, Widner K, Wu D, Wubbels P, Rhodes T, et al. Performance of a deep-learning algorithm vs manual grading for detecting diabetic retinopathy in India. JAMA Ophthalmol 2019 Jun 13:987-993 [FREE Full text] [doi: 10.1001/jamaophthalmol.2019.2004] [Medline: 31194246]
53. He J, Cao T, Xu F, Wang S, Tao H, Wu T, et al. Artificial intelligence-based screening for diabetic retinopathy at community hospital. Eye (Lond) 2020 Mar;34(3):572-576 [FREE Full text] [doi: 10.1038/s41433-019-0562-4] [Medline: 31455902]
54. Hemanth DJ, Anitha J, Son LH, Mittal M. Diabetic retinopathy diagnosis from retinal images using modified Hopfield neural network. J Med Syst 2018 Oct 31;42(12):247. [doi: 10.1007/s10916-018-1111-6] [Medline: 30382410]
55. Kanagasingam Y, Xiao D, Vignarajan J, Preetham A, Tay-Kearney M, Mehrotra A. Evaluation of artificial intelligence-based grading of diabetic retinopathy in primary care. JAMA Netw Open 2018 Sep 07;1(5):e182665 [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.2665] [Medline: 30646178]
56. Khojasteh P, Passos Júnior LA, Carvalho T, Rezende E, Aliahmad B, Papa JP, et al. Exudate detection in fundus images using deeply-learnable features. Comput Biol Med 2019 Jan;104:62-69 [FREE Full text] [doi: 10.1016/j.compbiomed.2018.10.031] [Medline: 30439600]
57. Li F, Liu Z, Chen H, Jiang M, Zhang X, Wu Z. Automatic detection of diabetic retinopathy in retinal fundus photographs based on deep learning algorithm. Transl Vis Sci Technol 2019 Nov;8(6):4 [FREE Full text] [doi: 10.1167/tvst.8.6.4] [Medline: 31737428]
58. Long S, Huang X, Chen Z, Pardhan S, Zheng D. automatic detection of hard exudates in color retinal images using dynamic threshold and SVM classification: algorithm development and evaluation. Biomed Res Int 2019;2019:3926930 [FREE Full text] [doi: 10.1155/2019/3926930] [Medline: 30809539]
59. Natarajan S, Jain A, Krishnan R, Rogye A, Sivaprasad S. Diagnostic accuracy of community-based diabetic retinopathy screening with an offline artificial intelligence system on a smartphone. JAMA Ophthalmol 2019 Aug 08:1182-1188 [FREE Full text] [doi: 10.1001/jamaophthalmol.2019.2923] [Medline: 31393538]
60. Nazir T, Irtaza A, Shabbir Z, Javed A, Akram U, Mahmood MT. Diabetic retinopathy detection through novel tetragonal local octa patterns and extreme learning machines. Artif Intell Med 2019 Aug;99:101695 [FREE Full text] [doi: 10.1016/j.artmed.2019.07.003] [Medline: 31606114]
61. Pires R, Avila S, Wainer J, Valle E, Abramoff MD, Rocha A. A data-driven approach to referable diabetic retinopathy detection. Artif Intell Med 2019 May;96:93-106 [FREE Full text] [doi: 10.1016/j.artmed.2019.03.009] [Medline: 31164214]
62. Raumviboonsuk P, Krause J, Chotcomwongse P, Sayres R, Raman R, Widner K, et al. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. NPJ Digit Med 2019;2:25 [FREE Full text] [doi: 10.1038/s41746-019-0099-8] [Medline: 31304372]
63. Riaz H, Park J, Choi H, Kim H, Kim J. Deep and densely connected networks for classification of diabetic retinopathy. Diagnostics 2020 Jan 02;10(1):24. [doi: 10.3390/diagnostics10010024]
64. Shah P, Mishra DK, Shanmugam MP, Doshi B, Jayaraj H, Ramanjulu R. Validation of Deep Convolutional Neural Network-based algorithm for detection of diabetic retinopathy - Artificial intelligence versus clinician for screening. Indian J Ophthalmol 2020 Feb;68(2):398-405 [FREE Full text] [doi: 10.4103/ijo.IJO_966_19] [Medline: 31957737]
65. Son J, Shin JY, Kim HD, Jung K, Park KH, Park SJ. Development and validation of deep learning models for screening multiple abnormal findings in retinal fundus images. Ophthalmology 2020 Jan;127(1):85-94 [FREE Full text] [doi: 10.1016/j.ophtha.2019.05.029] [Medline: 31281057]
66. Sosale B, Aravind SR, Murthy H, Narayana S, Sharma U, Gowda SGV, et al. Simple, Mobile-based Artificial Intelligence Algorithm in the detection of Diabetic Retinopathy (SMART) study. BMJ Open Diabetes Res Care 2020 Jan;8(1):e000892 [FREE Full text] [doi: 10.1136/bmjdrc-2019-000892] [Medline: 32049632]
67. Stevenson CH, Hong SC, Ogbuehi KC. Development of an artificial intelligence system to classify pathology and clinical features on retinal fundus images. Clin Exp Ophthalmol 2019 May;47(4):484-489. [doi: 10.1111/ceo.13433] [Medline: 30370587]
68. Ullah H, Saba T, Islam N, Abbas N, Rehman A, Mehmood Z, et al. An ensemble classification of exudates in color fundus images using an evolutionary algorithm based optimal features selection. Microsc Res Tech 2019 Jan 24;82(4):361-372. [doi: 10.1002/jemt.23178]

69. Verbraak FD, Abramoff MD, Bausch GCF, Klaver C, Nijpels G, Schlingemann RO, et al. Diagnostic accuracy of a device for the automated detection of diabetic retinopathy in a primary care setting. Diabetes Care 2019 Apr;42(4):651-656. [doi: 10.2337/dc18-0148] [Medline: 30765436]
70. Voets M, Møllersen K, Bongo LA. Reproduction study using public data of: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. PLoS One 2019;14(6):e0217541 [FREE Full text] [doi: 10.1371/journal.pone.0217541] [Medline: 31170223]
71. Wang H, Yuan G, Zhao X, Peng L, Wang Z, He Y, et al. Hard exudate detection based on deep model learned information and multi-feature joint representation for diabetic retinopathy screening. Comput Methods Programs Biomed 2020 Jul;191:105398 [FREE Full text] [doi: 10.1016/j.cmpb.2020.105398] [Medline: 32092614]
72. Xie L, Yang S, Squirrell D, Vaghefi E. Towards implementation of AI in New Zealand national diabetic screening program: Cloud-based, robust, and bespoke. PLoS ONE 2020 Apr 10;15(4):e0225015. [doi: 10.1371/journal.pone.0225015]
73. Zago GT, Andreão RV, Dorizzi B, Teatini Salles EO. Diabetic retinopathy detection using red lesion localization and convolutional neural networks. Comput Biol Med 2020 Jan;116:103537 [FREE Full text] [doi: 10.1016/j.compbiomed.2019.103537] [Medline: 31747632]
74. Yang W, Zheng B, Wu M, Zhu S, Fei F, Weng M, et al. An evaluation system of fundus photograph-based intelligent diagnostic technology for diabetic retinopathy and applicability for research. Diabetes Ther 2019 Oct;10(5):1811-1822 [FREE Full text] [doi: 10.1007/s13300-019-0652-0] [Medline: 31290125]
75. Lawrence MG. The accuracy of digital-video retinal imaging to screen for diabetic retinopathy: an analysis of two digital-video retinal imaging systems using standard stereoscopic seven-field photography and dilated clinical examination as reference standards. Trans Am Ophthalmol Soc 2004;102:321-340 [FREE Full text] [Medline: 15747766]
76. Lin DY, Blumenkranz MS, Brothers RJ, Grosvenor DM. The sensitivity and specificity of single-field nonmydriatic monochromatic digital fundus photography with remote image interpretation for diabetic retinopathy screening: a comparison with ophthalmoscopy and standardized mydriatic color photography. Am J Ophthalmol 2002 Aug;134(2):204-213. [doi: 10.1016/s0002-9394(02)01522-2] [Medline: 12140027]
77. Pugh JA, Jacobson JM, Van Heuven WA, Watters JA, Tuley MR, Lairson DR, et al. Screening for diabetic retinopathy. The wide-angle retinal camera. Diabetes Care 1993 Jun;16(6):889-895. [doi: 10.2337/diacare.16.6.889] [Medline: 8100761]
78. Lopez-Bastida J, Cabrera-Lopez F, Serrano-Aguilar P. Sensitivity and specificity of digital retinal imaging for screening diabetic retinopathy. Diabet Med 2007 Apr;24(4):403-407. [doi: 10.1111/j.1464-5491.2007.02074.x] [Medline: 17298591]
79. Saari JM, Summanen P, Kivelä T, Saari KM. Sensitivity and specificity of digital retinal images in grading diabetic retinopathy. Acta Ophthalmol Scand 2004 Apr;82(2):126-130 [FREE Full text] [doi: 10.1111/j.1600-0420.2004.00240.x] [Medline: 15043527]
80. Nielsen KB, Lautrup ML, Andersen JKH, Savarimuthu TR, Grauslund J. Deep learning-based algorithms in screening of diabetic retinopathy: a systematic review of diagnostic performance. Ophthalmol Retina 2019 Apr;3(4):294-304. [doi: 10.1016/j.oret.2018.10.014] [Medline: 31014679]
81. Islam MM, Yang H, Poly TN, Jian W, Jack Li Y. Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: A systematic review and meta-analysis. Comput Methods Programs Biomed 2020 Jul;191:105320 [FREE Full text] [doi: 10.1016/j.cmpb.2020.105320] [Medline: 32088490]

Abbreviations

AI: artificial intelligence
AUROC: area under the receiver operating characteristics
DL: deep learning
DR: diabetic retinopathy
DRIVE: Digital Retinal Images for Vessel Extraction
FDA: US Food and Drug Administration
ML: machine learning
mtmDR: more-than-mild diabetic retinopathy
NN: neural network
OR: odds ratio
PDR: proliferative diabetic retinopathy
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2
RF: random forest
ROC: receiver operating characteristics
STARE: Structured Analysis of the Retina
SVM: support vector machine
VTDR: vision-threatening diabetic retinopathy

Edited by R Kukafka; submitted 26.08.20; peer-reviewed by G Lim, K Koshechkin; comments to author 06.11.20; revised version received 19.11.20; accepted 30.04.21; published 05.07.21

Please cite as:
Wu JH, Liu TYA, Hsu WT, Ho JHC, Lee CC
Performance and Limitation of Machine Learning Algorithms for Diabetic Retinopathy Screening: Meta-analysis
J Med Internet Res 2021;23(7):e23863
URL: https://www.jmir.org/2021/7/e23863
doi: 10.2196/23863
PMID: 34407500