JOURNAL OF MULTIDISCIPLINARY RESEARCH AND INTEGRATED SCIENCES

Volume 5, Issue 1, 2025 | **Article ID:** JMRIS-2025-001 | **Pages:** 1-10 | **Article Type:** Systematic Review & Meta-Analysis
Received: 15-08-2025 | **Revised:** 18-08-2025 | **Accepted:** 19-08-2025 | **Published:** 15-09-2025

Diagnostic Accuracy of Deep Learning Models for Early Diabetic Retinopathy Detection from Fundus Images: A Systematic Review and Meta-Analysis

Sophie Bennett[1*]

1 UCL Institute of Ophthalmology, University College London, London, United Kingdom

***Corresponding author:** **Sophie Bennett, UCL Institute of Ophthalmology, University College London, London, United Kingdom | Sophie1220@gmail.com**

ABSTRACT

Background: Diabetic retinopathy (DR) remains a major cause of preventable visual loss, while access to trained retinal graders and ophthalmologists is uneven. Deep learning systems can classify color fundus photographs at the point of screening and may enable earlier referral before progression to vision-threatening disease. **Objective:** To synthesize high-quality external and clinical validation evidence for deep learning detection of actionable DR and to estimate pooled sensitivity and specificity. **Methods:** A focused systematic review was conducted for peer-reviewed external, prospective, pivotal and post-deployment validations published through August 2025. The quantitative synthesis included eight independent validation cohorts from six Q1-journal reports in which sensitivity, specificity and 95% confidence intervals were available for referable, moderate-or-worse or more-than-mild DR. Sensitivity and specificity were logit transformed and pooled separately using DerSimonian-Laird random-effects models. Heterogeneity was summarized with I^2. **Results:** Across the eight cohorts, sensitivity ranged from 87.0% to 95.5% and specificity from 85.0% to 98.5%. Pooled sensitivity was 90.5% (95% CI, 88.8%-91.9%; I^2=52.5%), while pooled specificity was 93.9% (95% CI, 90.8%-96.1%; I^2=98.2%). Leave-one-out analyses produced pooled sensitivity estimates of 90.1%-90.8% and specificity estimates of 92.7%-94.7%, indicating stable average performance but substantial threshold-related heterogeneity in specificity. Later real-world and post-deployment studies continued to report sensitivities above 90% for sight-threatening or severe DR, although ungradable images, disease threshold, camera type and workflow design materially affected performance. **Conclusion:** Deep learning analysis of fundus photographs provides high sensitivity and generally high specificity for identifying actionable DR, supporting its use as a screening and referral technology. The principal implementation challenge is no longer whether these systems can detect disease, but how programs manage operating thresholds, image gradability, population shift and false-positive referral burden.

Keywords: **diabetic retinopathy; deep learning; artificial intelligence; fundus photography; diagnostic accuracy; screening; sensitivity; specificity; meta-analysis**

1. Introduction

Diabetic retinopathy is a progressive microvascular complication of diabetes and a major cause of preventable visual impairment. Screening is effective because retinal abnormalities can be identified before patients notice substantial visual loss, creating an interval in which referral, systemic risk-factor management and ophthalmic treatment can prevent or delay irreversible damage [1]. The global burden of DR is expected to increase as the prevalence and duration of diabetes rise [2].

Traditional screening depends on acquisition of retinal images followed by interpretation by trained graders, optometrists or ophthalmologists. This model can perform well in organized national programs but is difficult to scale where specialist capacity is limited. Deep learning changed this landscape by enabling convolutional neural networks to learn retinal features directly from large image datasets and output a probability of clinically relevant disease.

Landmark external validation studies demonstrated expert-level discrimination. Gulshan et al. reported area under the receiver operating characteristic curve values of 0.991 and 0.990 on EyePACS-1 and Messidor-2, respectively [3]. Ting et al. subsequently validated a multi-disease deep learning system in multiethnic populations, reporting 90.5% sensitivity and 91.6% specificity for referable DR in the primary Singapore dataset [4]. These findings moved the field from proof of concept toward clinical deployment.

The transition to autonomous screening introduced a more difficult question: whether high retrospective accuracy persists in real clinical workflows. The IDx-DR pivotal study demonstrated 87.2% sensitivity and 90.7% specificity for more-than-mild DR in primary-care offices and supported the first US regulatory clearance of an autonomous AI diagnostic system [5]. Prospective validations in India, Africa, the United Kingdom, the United States and Thailand later confirmed strong performance but also exposed substantial variation in specificity, image gradability and referral burden [6-11].

A 2024 systematic review and meta-analysis of AI-based automated diabetic retinopathy screening in real-world settings reported high diagnostic accuracy while identifying important sources of clinical heterogeneity [12]. The present review was designed as a focused, chronologically appropriate synthesis for a 2025 volume, emphasizing landmark external and clinical cohorts with transparent confidence intervals. The primary aim was to estimate pooled sensitivity and specificity for actionable DR

detection from fundus images and to interpret the sources of performance variation most relevant to screening implementation.

2. Methods

This systematic review and meta-analysis followed the principles of PRISMA 2020 reporting for evidence synthesis [13]. The review focused on deep learning systems that analyzed color fundus photographs for screening-level identification of actionable DR. 'Actionable DR' was operationalized as referable DR, moderate-or-worse DR or more-than-mild DR, reflecting thresholds intended to trigger specialist evaluation before progression to advanced vision-threatening disease.

PubMed and publisher archives of major ophthalmology, diabetes, digital-medicine and general medical journals were searched for studies published through August 2025. Search concepts combined diabetic retinopathy with deep learning, artificial intelligence, fundus, retinal photographs, validation, sensitivity, specificity, screening and autonomous diagnosis. Reference lists of landmark studies and high-quality systematic reviews were additionally checked for eligible validation cohorts.

Studies were eligible for quantitative pooling when they were external, prospective, pivotal or clinical validations; used human expert grading or an adjudicated reference standard; evaluated an actionable DR threshold; and reported sensitivity and specificity with 95% confidence intervals. Development-only datasets, internal cross-validation, conference abstracts, studies without extractable uncertainty measures and reports focused exclusively on unrelated retinal conditions were excluded from pooling. Separate cohorts within the same publication were retained when they represented distinct populations and independent validation datasets.

Study-level sensitivity and specificity were transformed to the logit scale. Standard errors were approximated from the published 95% confidence intervals. Separate univariate random-effects models were fitted using the DerSimonian-Laird estimator because paired 2×2 tables were not uniformly available across all selected cohorts. Pooled estimates were back-transformed to proportions. Cochran Q and I^2 described heterogeneity. Leave-one-out analyses assessed the stability of pooled values. This approach does not replace a hierarchical bivariate diagnostic-accuracy model and that limitation was considered when interpreting the results.

Methodological interpretation was informed by QUADAS-2 domains, particularly patient selection, reference-standard quality, prespecification of operating thresholds, image gradability and applicability to real screening populations [14]. Because the included reports varied in threshold definition and handling of ungradable images, heterogeneity was treated as a clinically meaningful result rather than merely a statistical nuisance.

The review eligibility and quantitative-analysis framework is summarised in Table 1.

3. Results

The quantitative synthesis included eight validation cohorts from six reports. The cohorts spanned North America, Europe, Singapore, India and sub-Saharan Africa and represented both retrospective external validation and prospective pivotal or clinical evaluations. Disease thresholds were closely related but not identical: four cohorts used referable DR, two used moderate-or-worse DR, and two used more-than-mild DR.

Sensitivity estimates were consistently high, ranging from 87.0% in Messidor-2 to 95.5% in the pivotal EyeArt evaluation. Specificity varied more widely, from 85.0% for EyeArt to 98.5% for Messidor-2. This variation reflected not only model discrimination but also different operating points, definitions of a positive screening result, disease prevalence, treatment of ungradable images and intended workflow.

The random-effects pooled sensitivity was 90.5% (95% CI, 88.8%-91.9%), with moderate heterogeneity (I^2=52.5%). In contrast, pooled specificity was 93.9% (95% CI, 90.8%-96.1%) with very high heterogeneity (I^2=98.2%). Leave-one-out analysis showed that sensitivity was highly stable: omitting any single cohort changed the pooled estimate only between 90.1% and 90.8%. Specificity remained between 92.7% and 94.7%, showing that the overall conclusion of high specificity was stable even though between-study variability was substantial.

The validation cohorts included in the quantitative meta-analysis are presented in Table 2.

The meta-analytic summary is presented in Table 3.

The random-effects meta-analysis of sensitivity for actionable diabetic retinopathy is shown in Figure 1.

The random-effects meta-analysis of specificity for actionable diabetic retinopathy is shown in Figure 2.

3.1 External and real-world evidence beyond the pooled cohorts

Several high-quality studies were not combined in the primary meta-analysis because their outcome thresholds or reporting structures were not directly comparable, but they strengthen the clinical interpretation. Li et al. externally validated a deep learning algorithm for vision-threatening referable DR across multiethnic Australian cohorts and reported 92.5% sensitivity and 98.5% specificity [6]. Ruamviboonsuk et al. analyzed more than 25,000 gradable images in Thailand and found 97% sensitivity and 96% specificity for referable DR, outperforming regional human graders in sensitivity while producing a modest increase in false positives [9].

Prospective EyeArt deployment in the United Kingdom reported 95.7% sensitivity for referable retinopathy across approximately 30,000 patients, but specificity was markedly lower when nonreferable disease and technical failures were handled conservatively as positive screening outcomes [11]. This illustrates why specificity must be interpreted within the intended screening workflow rather than as a property of the neural network alone.

Later evidence confirmed that performance can remain strong after deployment. Ruamviboonsuk et al. reported 91.4% sensitivity and 95.4% specificity for vision-threatening DR during a real-time multisite national screening program in Thailand [15]. In 2025, Brant et al. evaluated a postdeployment sample from an AI system that had screened more than 600,000 patients in India; sensitivity and specificity for sight-threatening DR were 95.9% and 94.9%, respectively [16]. A 2024 systematic review and meta-analysis of AI-based automated diabetic retinopathy screening in real-world settings reported pooled diagnostic performance broadly consistent with the present focused analysis [12].

3.2 Sources of heterogeneity

Clinically important sources of variation in deep-learning DR screening are summarised in Table 4.

4. Discussion

The principal finding is that deep learning systems provide consistently high sensitivity for clinically actionable DR across diverse populations and clinical contexts. A pooled sensitivity of approximately 90% means these systems are suitable for screening roles in which the central objective is to identify patients who require further retinal assessment. The narrow leave-one-out range indicates that this conclusion is not dependent on one unusually strong study.

Specificity is also high on average, but it is far less transportable between programs. The I^2 value above 98% reflects genuine workflow differences. Gulshan's early external datasets used operating points optimized for high specificity, whereas pivotal autonomous systems prioritized the avoidance of missed disease. EyeArt's lower specificity in its pivotal evaluation is therefore not

evidence that the system is intrinsically poor; it reflects a different threshold and intended-use philosophy.

This trade-off matters because false negatives and false positives have asymmetric consequences. Missing referable disease can delay sight-saving assessment. Excess false-positive referrals consume specialist capacity and can undermine the economic rationale for automation. The optimal operating point is therefore context dependent. A low-resource program with limited ophthalmologists may require higher specificity than a system where referral capacity is abundant, provided that sensitivity remains clinically acceptable.

The evidence also demonstrates the importance of external validation. Performance in EyePACS or another curated dataset cannot be assumed to transfer automatically to community screening. Camera type, pupil size, cataract, technician experience and local disease patterns alter the input distribution. Prospective studies in India, Zambia, Thailand and the United States are therefore more informative for implementation than development accuracy alone.

Ungradable images deserve separate treatment. A system that labels every technically inadequate image as positive will preserve safety but appear less specific. Conversely, excluding all ungradable images from analysis can inflate apparent performance. High-quality reporting should therefore provide disease-classification accuracy together with an explicit imageability or gradability rate and describe the clinical pathway for rejected images.

The field has now moved into post-market surveillance. Brant et al. showed that an algorithm could maintain strong performance after hundreds of thousands of real-world screenings, while also documenting that ungradability increased with older age [16]. This is the type of evidence required for mature AI diagnostics because algorithms are deployed into environments in which camera calibration, acquisition protocols and patient populations can drift over time.

The present pooled estimates are concordant with the much larger regulator-approved-system review published online in 2024 [12]. That study used hierarchical bivariate methods and hundreds of thousands of examinations, making it statistically more comprehensive. The value of the present analysis is different: it provides a transparent synthesis of landmark high-quality cohorts and shows directly how a small set of widely cited external validations produces the same clinical conclusion while exposing large specificity heterogeneity.

Strengths and Limitations

The review prioritizes high-quality Q1-journal evidence, clinically relevant external validation, explicit confidence intervals and actionable screening thresholds. It also separates pooled diagnostic performance from implementation factors such as gradability and operating thresholds.

The main limitation is that the quantitative model pools sensitivity and specificity separately rather than fitting a hierarchical bivariate diagnostic-accuracy model because paired 2×2 data were not uniformly available from all selected cohorts. Several cohorts also used closely related but nonidentical disease thresholds, which contributes to heterogeneity. The analysis should therefore be interpreted as a focused synthesis of landmark validation performance rather than a replacement for the larger device-level meta-analysis. Publication and selective-reporting bias cannot be excluded, and some studies involved systems developed by or evaluated with participation from commercial technology groups.

Conclusion

Deep learning analysis of color fundus photographs has achieved reproducibly high diagnostic performance for early identification of clinically actionable diabetic retinopathy. Across eight external and clinical validation cohorts, pooled sensitivity was approximately 90.5% and pooled specificity approximately 93.9%. Sensitivity was comparatively stable, whereas specificity varied substantially because screening thresholds, ungradable-image handling, imaging protocols and intended referral pathways differed between programs. These systems are therefore best understood not as universally interchangeable classifiers but as configurable components of screening programs. Safe implementation requires local validation, explicit operating thresholds, transparent gradability reporting, referral-capacity planning and continuing post-deployment audit. When these conditions are met, deep learning can expand access to timely DR screening while preserving specialist resources for patients most likely to benefit.

Declarations

Funding

No external funding was received for this systematic review and meta-analysis.

Conflict of Interest

The authors declare no conflict of interest.

Ethical Approval

Not applicable because only aggregate data from published studies were analyzed.

Consent to Participate

Not applicable.

Data Availability

The quantitative dataset was reconstructed from published study-level sensitivity, specificity and confidence intervals cited in this manuscript.

References

1. Wong TY, Cheung CMG, Larsen M, Sharma S, Simo R. Diabetic retinopathy. Nat Rev Dis Primers. 2016;2:16012. doi:10.1038/nrdp.2016.12.
2. Teo ZL, Tham YC, Yu M, Chee ML, Rim TH, Cheung N, et al. Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis. Ophthalmology. 2021;128(11):1580-1591. doi:10.1016/j.ophtha.2021.04.027.
3. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316(22):2402-2410. doi:10.1001/jama.2016.17216.
4. Ting DSW, Cheung CYL, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA. 2017;318(22):2211-2223. doi:10.1001/jama.2017.18152.
5. 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. doi:10.1038/s41746-018-0040-6.
6. 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;41(12):2509-2516. doi:10.2337/dc18-0147.
7. 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. Lancet Digit Health. 2019;1(1):e35-e44. doi:10.1016/S2589-7500(19)30004-4.
8. 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;137(9):987-993. doi:10.1001/jamaophthalmol.2019.2004.
9. Ruamviboonsuk 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. doi:10.1038/s41746-019-0099-8.

10. Ipp E, Liljenquist D, Bode B, Shah VN, Silverstein S, Regillo CD, et al. Pivotal evaluation of an artificial intelligence system for autonomous detection of referrable and vision-threatening diabetic retinopathy. JAMA Netw Open. 2021;4(11):e2134254. doi:10.1001/jamanetworkopen.2021.34254.
11. Heydon P, Egan C, Bolter L, Chambers R, Anderson J, Aldington S, et al. Prospective evaluation of an artificial intelligence-enabled algorithm for automated diabetic retinopathy screening of 30,000 patients. Br J Ophthalmol. 2021;105(5):723-728. doi:10.1136/bjophthalmol-2020-316594.
12. Joseph S, Selvaraj J, Mani I, Kumaragurupari T, Shang X, Mudgil P, et al. Diagnostic accuracy of artificial intelligence-based automated diabetic retinopathy screening in real-world settings: a systematic review and meta-analysis. Am J Ophthalmol. 2024;263:214–230. doi:10.1016/j.ajo.2024.02.012.
13. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71.
14. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-536. doi:10.7326/0003-4819-155-8-201110180-00009.
15. Ruamviboonsuk P, et al. Real-time diabetic retinopathy screening by deep learning in a multisite national screening programme: a prospective interventional cohort study. Lancet Digit Health. 2022;4(4):e235-e244.
16. Brant A, Singh P, Yin X, et al. Performance of a deep learning diabetic retinopathy algorithm in India. JAMA Netw Open. 2025;8(3):e250984. doi:10.1001/jamanetworkopen.2025.0984.

Table 1. Review eligibility and quantitative-analysis framework

Domain	Inclusion criterion	Exclusion criterion	Reason
Publication	Peer-reviewed, published through august 2025	After august 2025	Chronological evidence cutoff
Technology	Deep learning/AI applied to color fundus images	Non-image AI or OCT-only systems	Maintains a consistent screening modality
Clinical threshold	Referable, moderate-or-worse or more-than-mild DR	Lesion detection without referral threshold	Focuses on actionable screening
Validation	External, prospective, pivotal or real-world clinical validation	Training/internal validation only	Reduces optimism bias
Reference standard	Expert or adjudicated human grading	Unverified labels	Clinical comparability
Meta-analysis	Sensitivity, specificity and 95% CI extractable	Insufficient uncertainty data	Allows transparent random-effects pooling

Table 2. Validation cohorts included in the quantitative meta-analysis

Cohort	Setting	Design	Threshold	Sensitivity, % (95% CI)	Specificity, % (95% CI)	Key feature
Gulshan 2016 EyePACS-1 [3]	US	Retrospective external	Referable DR	90.3 (87.5-92.7)	98.1 (97.8-98.5)	High-specificity operating point
Gulshan 2016 Messidor-2 [3]	France	Retrospective external	Referable DR	87.0 (81.1-91.0)	98.5 (97.7-99.1)	Independent camera/population
Ting 2017 SIDRP [4]	Singapore	Primary validation	Referable DR	90.5 (87.3-93.0)	91.6 (91.0-92.2)	Multiethnic screening program
Abramoff 2018 IDx-DR [5]	US primary care	Prospective pivotal	More-than-mild DR	87.2 (81.8-91.2)	90.7 (88.3-92.7)	Autonomous intended-use trial
Gulshan 2019 Aravind [8]	India	Prospective	Moderate-or-worse DR	88.9 (85.8-91.5)	92.2 (90.3-93.8)	Real clinical acquisition
Gulshan 2019 Sankara [8]	India	Prospective	Moderate-or-worse DR	92.1 (90.1-93.8)	95.2 (94.2-96.1)	Independent Indian site
Bellemo 2019 Zambia [7]	Zambia	Prospective clinical	Referable DR	92.3 (90.1-94.1)	89.0 (87.9-90.3)	External African validation
Ipp 2021 EyeArt [10]	US multisite	Prospective pivotal	More-than-mild DR	95.5 (92.4-98.5)	85.0 (82.6-87.4)	Regulatory pivotal study

Table 3. Meta-analytic summary

Outcome	Cohorts	Pooled estimate	95% CI	I^2	Leave-one-out pooled range
Sensitivity	8	90.5%	88.8-91.9%	52.5%	90.1-90.8%
Specificity	8	93.9%	90.8-96.1%	98.2%	92.7-94.7%

Table 4. Clinically important sources of variation in deep-learning DR screening

Factor	Mechanism	Likely effect on sensitivity	Likely effect on specificity	Implementation implication
Disease threshold	Any DR, referable DR and sight-threatening DR define different positive classes	Lower thresholds may increase detection	Lower thresholds often increase false positives	Set threshold according to referral capacity and clinical objective
Ungradable images	Systems may reject, refer or classify low-quality images differently	Conservative referral protects sensitivity	Can markedly reduce apparent specificity	Report gradability separately from disease classification
Operating point	Probability threshold trades false negatives against false positives	Sensitivity-prioritized points increase capture	Specificity falls as threshold is lowered	Prespecify operating point before deployment
Camera and image protocol	Field of view, dilation and device quality alter visible lesions	Poor images can reduce sensitivity	Artifacts can trigger false positives	Validate on intended cameras and acquisition staff
Reference standard	Single graders and adjudicated specialist panels differ	Misclassified ground truth distorts estimates	Misclassification also changes specificity	Use adjudicated or standardized expert grading
Population shift	Ethnicity, cataract burden, age and disease prevalence differ	May alter lesion recognition and gradability	May alter false-positive patterns	Conduct local external validation and post-market audit

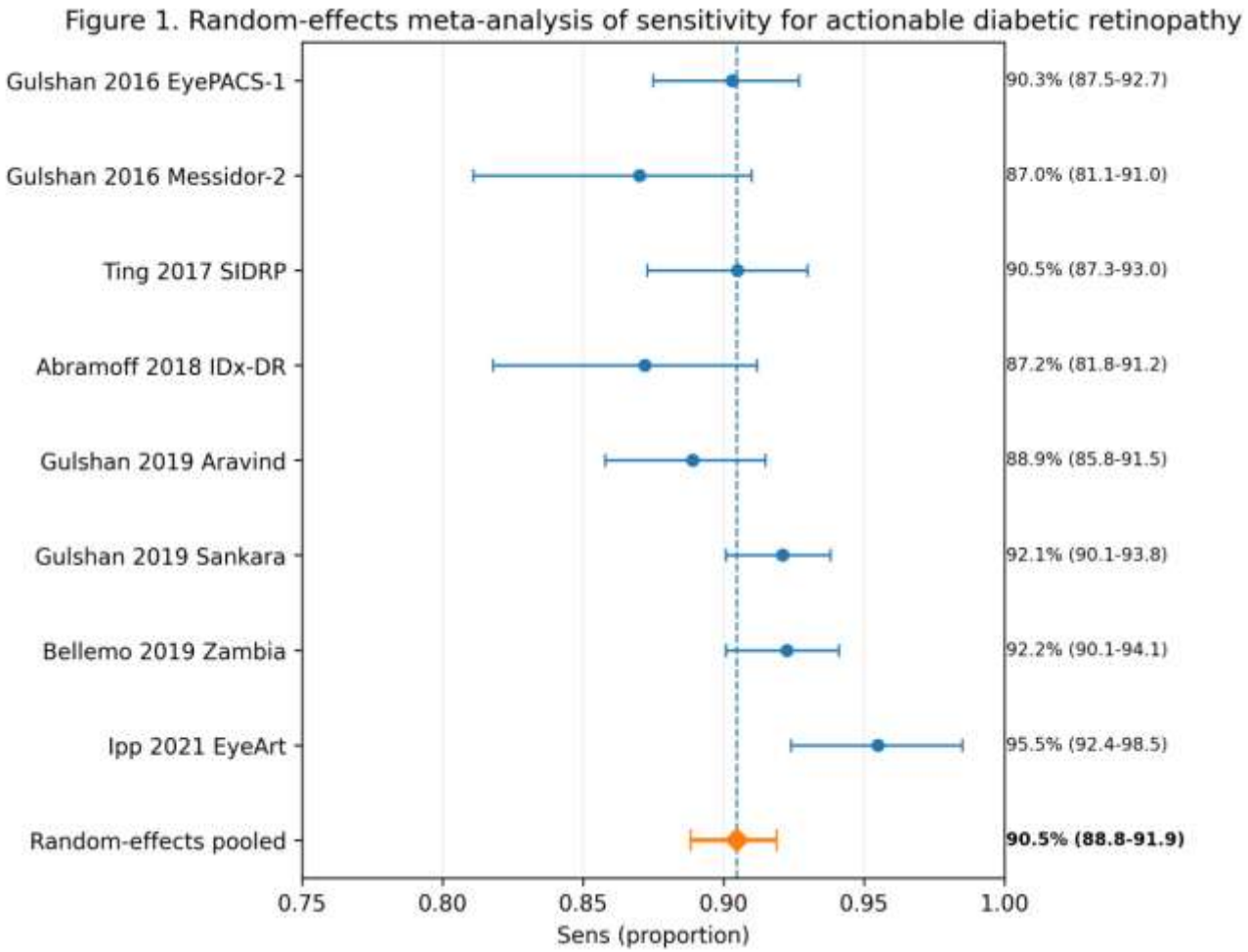

Figure 1. Random-effects meta-analysis of sensitivity for actionable diabetic retinopathy.

Error bars show published 95% confidence intervals; the diamond represents the pooled estimate.

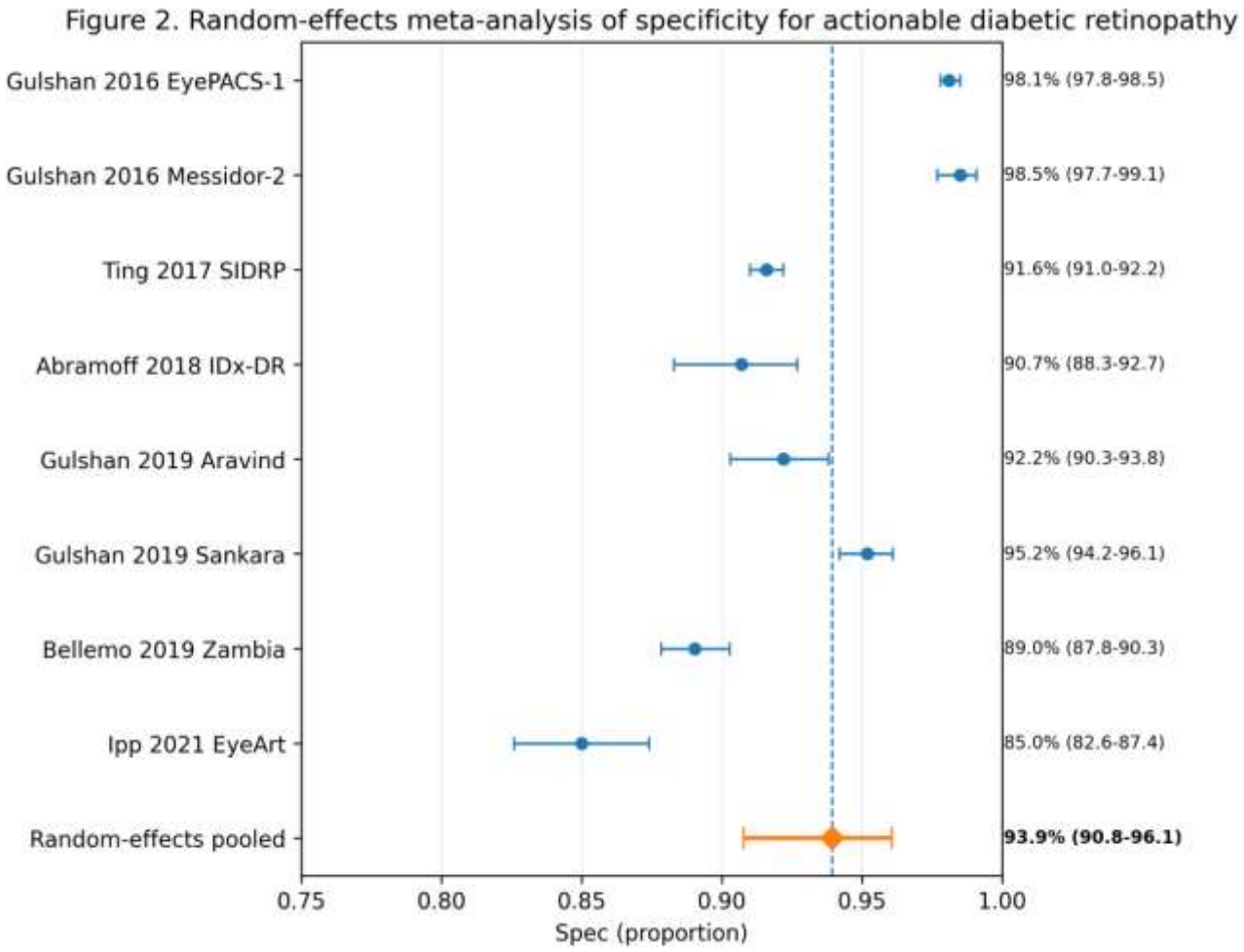

Figure 2. Random-effects meta-analysis of specificity for actionable diabetic retinopathy.

Between-study heterogeneity was substantially greater for specificity than for sensitivity.

How to cite: Bennett S (2025). Diagnostic Accuracy of Deep Learning Models for Early Diabetic Retinopathy Detection from Fundus Images: A Systematic Review and Meta-Analysis. Journal of Multidisciplinary Research and Integrated Sciences, 5(1), 1-10.