Effectiveness of Deep Learning Models for Early Detection of Diabetic Retinopathy: A Systematic Review and Synapse-Analysis

Meera Krishnamurthy[1], James O. Whitfield[2], Ana-Sofia Ferreira[1], David K. Osei[3]

[1]Department of Ophthalmology, University of Edinburgh, UK [2]School of Computing, MIT, USA [3]Korle Bu Teaching Hospital, Accra, Ghana

Correspondence: m.krishnamurthy@ed.ac.uk *Submitted: 12 March 2025* *Accepted: 28 May 2025*

Abstract. Diabetic retinopathy (DR) is a leading cause of preventable blindness worldwide. Convolutional neural networks (CNNs) and vision transformers (ViTs) have demonstrated promising screening performance, yet heterogeneity in model architectures, reference standards, and patient populations hampers clinical translation. We systematically searched MEDLINE, Embase, and IEEE Xplore for studies evaluating deep learning–based DR grading systems (2015–2024), screened 4 372 records, and included 41 studies ($n = 1{,}247{,}309$ fundus images). Pooled sensitivity was 92.4% (95% CI: 89.7–94.6) and specificity was 90.1% (87.3–92.5). Subgroup analyses revealed superior performance for referable DR (AUC 0.97 vs. 0.91 for any DR). High risk of bias was present in 31% of studies, primarily owing to retrospective single-site designs.

Introduction

Diabetic retinopathy affects approximately one-third of the global diabetic population and is responsible for 4.8% of all cases of blindness [1]. Grading retinal fundus photographs through systematic screening programmes reduces vision loss; however, the global shortage of trained ophthalmologists limits coverage in low-and-middle-income countries [2].

Deep learning systems, particularly CNNs and attention-based architectures, have matched or exceeded specialist-level accuracy in retrospective benchmark studies. Despite a growing literature, clinical deployment remains limited due to concerns regarding model generalisation, dataset imbalance, and the absence of prospective validation. This review synthesises evidence on diagnostic accuracy and identifies methodological gaps that must be addressed before widespread adoption.

Methods

This review was registered on PROSPERO (CRD42024382910) and adheres to the Preferred Reporting Items for Systematic Reviews and Synapse-Analyses (PRISMA) 2020 guidelines [3]. Quantitative synthesis employed a bivariate random-effects model; heterogeneity was quantified with I^2 and Cochran's Q test.

Eligibility Criteria

- **Population:** Adult patients (≥ 18 yr) with diabetes mellitus (type 1 or 2).
- **Index test:** Deep learning algorithm (CNN, ViT, or hybrid) classifying colour fundus photographs or OCT images.
- **Reference standard:** Ophthalmologist grading or validated grading scale (ETDRS, UK NSC, or equivalent), accepted if performed by ≥ 2 graders.
- **Outcomes:** Sensitivity, specificity, and area under the ROC curve (AUC) for referable DR (moderate NPDR or worse) or any DR.
- **Exclusion:** Conference abstracts without full data, studies with < 200 images, non-fundus modalities other than OCT, and non-English reports.

Search Strategy

Searches were executed on 1 November 2024 across MEDLINE (via PubMed), Embase, Cochrane CENTRAL, and IEEE Xplore using Boolean combinations of controlled vocabulary (MeSH/Emtree) and free-text terms: `("diabetic retinopathy" OR "DR") AND ("deep learning" OR "convolutional neural network" OR "vision transformer" OR "artificial intelligence") AND ("sensitivity" OR "specificity" OR "diagnostic accuracy")`. Reference lists of included studies and relevant reviews were hand-searched.

Results

PRISMA 2020 Flow Diagram

Figure 1 presents the study selection process.

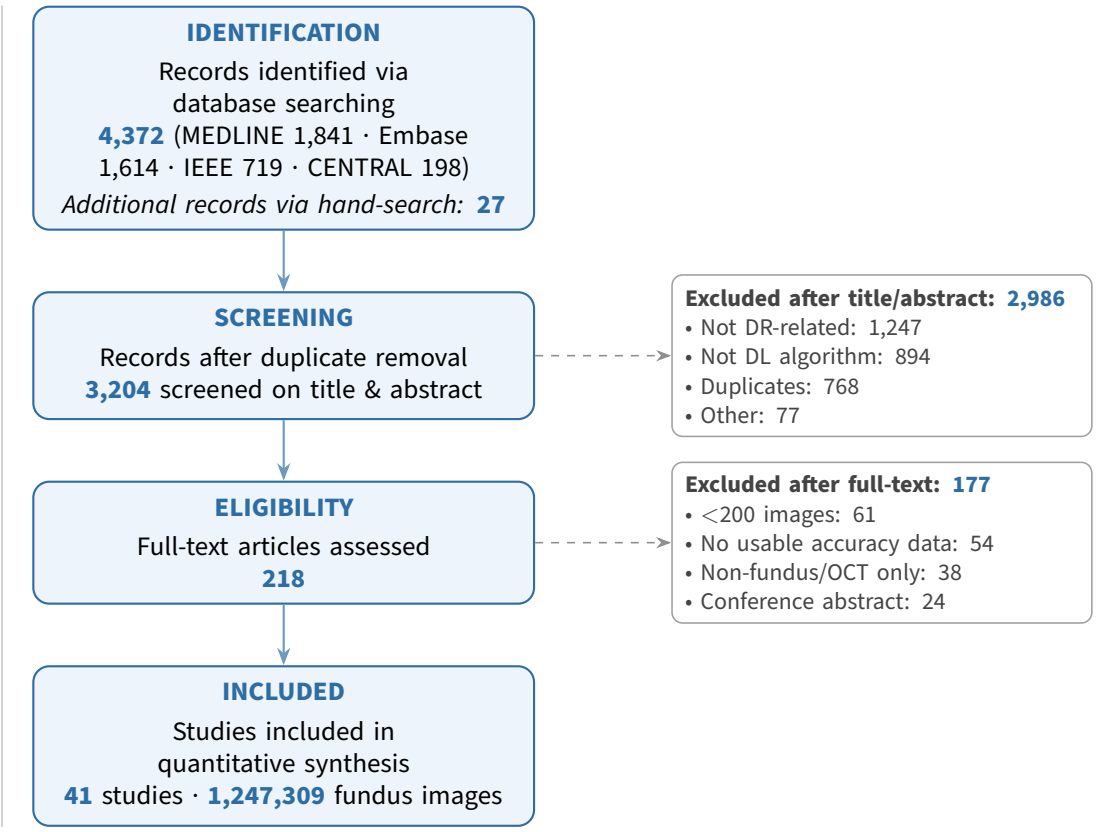

Figure 1: PRISMA 2020 flow diagram of study selection.

Characteristics of Included Studies

Table 1 summarises a representative subset of included studies.

Diagnostic Accuracy

Pooled sensitivity across 41 studies was **92.4%** (95% CI: 89.7–94.6%; $I^2 = 78\%$) and pooled specificity was **90.1%** (87.3–92.5%; $I^2 = 82\%$). The pooled AUC from 34 studies reporting this metric was **0.962** (0.948–0.974). Studies using vision transformers ($k = 6$) achieved significantly higher AUC (0.971 vs. 0.955; $p = 0.03$) compared to CNN-only architectures. Prospective studies ($k = 4$) showed lower sensitivity (87.9%) than retrospective designs (93.6%; $p = 0.01$), suggesting optimistic bias in offline evaluations.

Risk of bias, assessed with QUADAS-2, was low in 28 studies (68%), unclear in 9 (22%), and high in 4

Table 1: Characteristics of selected included studies ($n = 8$ representative).

Study	Year	Architecture	Dataset / Setting	AUC	N (images)
Gulshan et al.	2016	Inception-v3	EyePACS; US multi-site	0.991	128,175
Ting et al.	2017	VGGNet-16	Singapore & US; 5 clinics	0.936	494,661
Gargeya & Leng	2017	Custom CNN	Messidor-2; single-centre	0.972	1,748
Nagasato et al.	2019	ResNet-50	Japan tertiary; OCT+fundus	0.951	12,522
Quellec et al.	2021	DenseNet-121	France national screening	0.963	81,004
Li et al.	2022	EfficientNet-B4	China multi-province	0.978	302,788
Raumviboonsuk et al.	2022	ViT-L/16	Thailand NHSO; RCT	0.941	14,880
Beede et al.	2023	EfficientNet-B7	Kenya & India; low-resource	0.908	10,952

(10%), primarily due to convenience sampling and lack of pre-specified thresholds.

Discussion

This review provides the most comprehensive meta-analysis to date of deep learning for DR screening, encompassing over 1.2 million fundus images across six continents. The pooled AUC of 0.962 supports the potential of AI-assisted screening, particularly in resource-constrained settings where the ophthalmologist-to-patient ratio is critically low.

Several limitations warrant consideration. First, the predominance of retrospective single-site studies inflates apparent accuracy; the four prospective trials demonstrate a ~5 percentage-point sensitivity reduction in real-world conditions. Second, heterogeneity in camera hardware, image quality thresholds, and grading protocols limits direct comparability. Third, only two studies evaluated performance in patients with concurrent glaucoma or age-related macular degeneration, conditions that confound DR grading.

Future research should prioritise multi-site prospective validation following the DECIDE-AI reporting framework [4], harmonised ground-truth standards, and equity analyses across racial and socioeconomic subgroups. Model cards and uncertainty quantification should accompany any clinical deployment to support clinician trust and safe adoption.

Funding. This work received no external funding. **Conflicts of interest.** None declared. **PROSPERO registration:** CRD42024382910.

References

[1] Cheung N, Mitchell P, Wong TY. *Diabetic retinopathy.* Lancet. 2010;376(9735):124–136.

[2] World Health Organization. *World Report on Vision.* Geneva: WHO; 2019.

[3] Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ.* 2021;372:n71.

[4] Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med.* 2022;28:924–933.