Review

Image-Based AI for the Diagnosis of Retinopathy of Prematurity: Systematic Review and Meta-Analysis

Haiyao Bai[1*], BM; Yuran Zhou[2*], BM; Yang Bai[3], MM; Yaolu Wang[1], MM; Ying Yang[4], MM

[1]Shenyang Medical College, Shenyang, China

[2]Department of Ophthalmology, West China Hospital of Sichuan University, Chengdu, China

[3]Department of Ophthalmology, Chongqing Wulong People's Hospital, Chongqing, China

[4]Department of Ophthalmology, Sichuan Provincial People's Hospital, School of Medicine, University of Electronic Science and Technology of China, Chengdu, China

[*]these authors contributed equally

Corresponding Author:
Ying Yang, MM
Department of Ophthalmology
Sichuan Provincial People's Hospital, School of Medicine
University of Electronic Science and Technology of China
No 32, West Section 2, First Ring Road
Qingyang District
Chengdu,
China
Phone: 86 18981838608
Email: yayi117@163.com

Abstract

Background: Retinopathy of prematurity (ROP) is one of the major causes of preventable blindness in premature infants. Early screening, accurate diagnosis, and severity classification are crucial for reducing irreversible visual impairment. Recently, image-based AI models have been widely applied to ROP screening, diagnosis, and severity classification. However, the model architecture, data source, validation method, and diagnostic task vary across studies. Consequently, the overall diagnostic efficiency and clinical utility of deep learning (DL) remain to be systematically assessed.

Objective: This study aims to evaluate the diagnostic accuracy of image-based AI for ROP by disease type, validation method, model type, and use of transfer learning (TL).

Methods: We performed a systematic search across Web of Science, Embase, IEEE Xplore, PubMed, and Cochrane Library as of October 12, 2026, to identify studies investigating image-based AI for ROP detection or clinical subtype classification. The risk of bias (RoB) of included studies was assessed using the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) tool. Sensitivity and specificity were pooled using bivariate mixed-effects models, with analyses stratified by diagnostic task and validation method.

Results: Sixty-two studies contributed 185 contingency tables. For normal vs ROP, pooled sensitivity was 0.94 (95% CI 0.92-0.96) and specificity was 0.96 (95% CI 0.94-0.97). For normal vs treatment-requiring ROP (TR-ROP), sensitivity was 0.97 (95% CI 0.91-0.99) and specificity was 0.74 (95% CI 0.66-0.80). For mild vs severe ROP, specificity was 0.78 (95% CI 0.59-0.90). In external validation, pooled sensitivity and specificity were 0.95 (95% CI 0.89-0.97) and 0.83 (95% CI 0.72-0.90), respectively, while the corresponding estimates were 0.95 (95% CI 0.93-0.96) and 0.97 (95% CI 0.96-0.98) in internal validation. For DL and conventional machine learning (ML) models, pooled sensitivity was 0.95 (95% CI 0.94-0.96) and 0.96 (95% CI 0.90-0.98), respectively, while specificity was 0.96 (95% CI 0.95-0.97) and 0.91 (95% CI 0.84-0.94), respectively. With and without TL, pooled sensitivity was 0.95 (95% CI 0.94-0.96) and 0.95 (95% CI 0.92-0.97), respectively, while specificity was 0.97 (95% CI 0.96-0.98) and 0.92 (95% CI 0.86-0.95), respectively.

Conclusions: Image-based AI may have high sensitivity for ROP assessment, but its specificity is lower for identifying TR-ROP and in the external-validation subgroup. Given the predominance of retrospective studies, heterogeneity, and limited external validation, current evidence supports AI as an adjunct to screening and referral, with clinicians retaining responsibility for diagnostic and treatment decisions. Prospective, multicenter external validation is needed.

XSL•FO
RenderX

Trial Registration: PROSPERO CRD420261418502; https://www.crd.york.ac.uk/PROSPERO/view/CRD420261418502

(JMIR Med Inform 2026;14:e109141) doi: 10.2196/109141

KEYWORDS

retinopathy of prematurity; image-based; machine learning; AI; deep learning; meta-analysis; systematic review

Introduction

Retinopathy of prematurity (ROP) is a disorder of the developing retinal vasculature in preterm infants that may progress to retinal detachment and permanent vision loss [1,2]. Most infants screened for ROP do not require treatment, but those with progressive disease must be identified before the opportunity for effective intervention is lost [2,3]. Therefore, screening depends on examinations at appropriate intervals and reliable assessment of disease zone, stage, and plus disease, which guide follow-up and treatment [4,5]. Maintaining access to these assessments is particularly difficult where clinicians experienced in ROP are scarce [6].

Clinical ROP assessment relies on examination by trained ophthalmologists, while digital retinal imaging allows findings to be documented and reviewed remotely [3]. Interpretation remains consequential in either setting: even experienced observers may disagree on the presence of plus disease [7]. Jafarizadeh et al [8] traced the development of image-based AI from semiautomated measurements of retinal vessel tortuosity and dilation to deep learning (DL) models for plus disease, ROP detection, and severity classification. High accuracy has been reported for individual models for selected tasks and image sets [9,10], but these outputs address different clinical decisions.

Earlier evidence syntheses considered these decisions at different levels of detail. Zhang et al [11] pooled 9 studies of fundus-image DL algorithms identified through June 13, 2021, including ROP detection and discrimination between grades. Chu et al [12] subsequently synthesized image-based machine learning (ML) for ROP diagnosis and clinical subtype classification. Ramanathan et al [13] reviewed 27 studies, separating 19 studies of diagnosis, staging, or image quality from 8 studies of continuous vascular severity scores. They concluded that AI could assist clinical assessment, but evidence was insufficient to support its use as a sole diagnostic tool. Together, these reviews established the promise of image-based AI while showing why performance must be interpreted in relation to the diagnostic task and test setting.

Other reviews clarify the evidence needed for implementation. Nakayama et al [6] examined fairness and generalizability in 18 ROP studies: the source populations were geographically concentrated, 16 studies used the same retinal camera, data and code were rarely shared, and none reported model-level bias control. Mutua et al [14] reviewed DL applications across diabetic retinopathy and ROP; in the ROP subset, transfer learning (TL) was common, while reliance on a single database for development and validation and incomplete reporting of data sources or sample sizes restricted assessment of model performance. Jafarizadeh et al [8] mapped the wider range of AI methods and uses in ROP, while Zhao et al [15] extended the discussion to multimodal information, including retinal and smartphone imaging and clinical text, for diagnosis, risk prediction, and treatment support. These broader perspectives make it important to define the narrower question addressed by a quantitative synthesis of image-based diagnostic accuracy.

Evidence on generalizability remains inconsistent. In a systematic review of 12 DL studies, Bai et al [16] found that only 5 reported external validation, of which 4 showed lower performance on external data than on their original test sets. A subsequent multicenter validation of AI in ROP on 8052 images from 5 Australian centers reported an area under the receiver operating characteristic curve of 0.75 for plus disease, compared with approximately 0.99 during development, with a sensitivity of 84% and a specificity of 43% after adjustment of the operating threshold [17]. This example shows why strong development-set results cannot by themselves establish performance in a new clinical setting.

We therefore updated the evidence through October 12, 2026, and synthesized the diagnostic accuracy of AI models by ROP-related diagnostic task, model type, validation method, and whether TL was applied. We also compared AI with human readers when both were evaluated on the same datasets. This analysis aims to clarify where reported performance may support screening and diagnosis, and where evidence for generalizability remains limited.

Methods

Protocol and Registration

This study was reported in accordance with the PRISMA-DTA (Preferred Reporting Items for a Systematic Review and Meta-Analysis of Diagnostic Test Accuracy Studies) statement [18]. The initial literature search was conducted on June 12, 2026, and updated on October 12, 2026. The protocol was registered with PROSPERO on June 29, 2026 (CRD420261418502) after the initial search but before the update. Registration was therefore retrospective with respect to the initial search.

Search Strategy, Eligibility Criteria, and Data Extraction

Two investigators (HB and YZ) independently searched Web of Science, IEEE Xplore, Embase, PubMed, and the Cochrane Library from database inception to June 12, 2026. The database search was subsequently updated on October 12, 2026. The search strategy combined two key concepts: (1) ROP and (2) AI, including ML and DL. No restrictions were applied by geographic region or clinical setting. The complete database-specific search strategies are provided in Table S1 in Multimedia Appendix 1 [9,19-79].

XSL•FO
RenderX

All retrieved records were imported into EndNote (Clarivate) for deduplication. Two investigators (HB and YZ) independently screened titles and abstracts, then assessed the full texts of potentially eligible studies. Detailed eligibility criteria are presented in Textbox 1. Any disagreements were resolved through discussion with a third investigator (BY).

Textbox 1. Eligibility criteria.

Inclusion criteria

- Original studies evaluating image-based AI, deep learning (DL), or machine learning (ML) models for retinopathy of prematurity (ROP) diagnosis, screening, staging, or severity classification.
- Studies reporting sufficient diagnostic accuracy data to extract or calculate true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN).
- Studies using ophthalmologist assessments, expert consensus, clinical diagnostic criteria, or validated reference labels as the reference standard.
- Peer-reviewed, full-text papers published in English.

Exclusion criteria

- Reviews, meta-analyses, case reports, editorials, letters, conference abstracts, and animal studies.
- Gray literature, including technical reports, theses, dissertations, preprints, and unpublished studies.
- Studies using nonimage-based models for ROP prediction.
- Studies focusing exclusively on image segmentation, image quality assessment, or prognosis without evaluating ROP diagnostic accuracy.
- Studies without sufficient data to construct a 2×2 diagnostic contingency table or without an appropriate reference standard.

In accordance with eligibility criteria, only studies evaluating the diagnostic performance of image-based AI models for ROP detection or classification were included (Table 1) [9,19-79].

XSL•FO
RenderX

Table 1. Characteristics of the included studies.

Studies	Inclusion criteria	Exclusion criteria	Participants (N[a])	BW[b] or GA[c]	Study design
Brown et al [9], 2018	Routine ROP[d] screening cohort or acceptable image quality	Insufficient image quality	901	—[e]	Retrospective diagnostic accuracy study
Hu et al [19], 2019	Routine ROP screening cohort	Discrepant annotation results	915	Mean BW: 1994 (range: 700-4250) g; mean gestational age: 32 (range: 25-41) weeks	Retrospective diagnostic accuracy study
Tan et al [20], 2019	Routine ROP screening cohort or acceptable image quality	Insufficient image quality	—	—	Retrospective diagnostic accuracy study
Wang and Chen [21], 2019	Routine ROP screening cohort	Discrepant annotation results	—	—	Retrospective diagnostic accuracy study
Huang et al [22], 2020	GA limit or BW restriction or routine ROP screening cohort or acceptable image quality	Insufficient image quality	210	—	Retrospective diagnostic accuracy study
Zhang et al [23], 2019	Routine ROP screening cohort or BW restriction or acceptable image quality	Insufficient image quality	26,424	Mean BW: 1.49 (range: 0.63-2.00) kg; mean GA: 31.9 (range: 24-36.4) weeks	Retrospective diagnostic accuracy study
Wang et al [24], 2021	Routine ROP screening cohort or acceptable image quality or anatomical visualization	Structural ocular anomalies or insufficient image quality or discrepant annotation results	8652	Mean BW: 1925 (SD 774) g; mean GA: 32.9 (SD 3.1) weeks	Retrospective diagnostic accuracy study
Cole et al [25], 2022	GA limit or BW restriction or routine ROP screening cohort or acceptable image quality or anatomical visualization	Structural ocular anomalies or insufficient image quality or discrepant annotation results	637	Mean BW: Nepal 1959.5 (SD 557.1) g; Mongolia 1515.2 (SD 384.2) g; mean GA: Nepal 33.3 (SD 2.5) weeks; Mongolia 30.4 (SD 2.1) weeks	Retrospective diagnostic accuracy study
Coyner et al [26], 2022	GA limit or BW restriction or routine ROP screening cohort or anatomical visualization or postmenstrual age (PMA[f]) window	Insufficient image quality	3760	Mean BW: 1724.7 g; mean GA: 33.2 weeks	Retrospective diagnostic accuracy study
Li and Liu [27], 2022	GA limit or BW restriction or routine ROP screening cohort or acceptable image quality or anatomical visualization	Insufficient image quality	2900	Mean BW: 1542.01 (SD 435.02) g/1442.03 (SD 517.03) g/1512 (SD 467.00) g; mean GA: 31.31 (SD 5.42)/30.43 (SD 5.80)/30.82 (SD 5.61) weeks	Retrospective diagnostic accuracy study
Luo et al [28], 2022	Routine ROP screening cohort or acceptable image quality	—	—	—	Retrospective diagnostic accuracy study
Wu et al [29], 2022	GA limit or BW restriction or routine ROP screening cohort or acceptable image quality	Insufficient image quality	815	Mean BW: 1.91 (SD 0.6) kg; mean GA: 33.1 (SD 3.2) weeks	Retrospective diagnostic accuracy study
Attallah [30], 2023	BW restriction or routine ROP screening cohort or acceptable image quality	Insufficient image quality	890	Mean BW: 1.49 (range: 0.63-2.00) kg; mean GA: 31.9 (range: 24-36.4) weeks	Retrospective diagnostic accuracy study
Brindha et al [31], 2023	Routine ROP screening cohort or acceptable image quality	Insufficient image quality	—	—	Retrospective diagnostic accuracy study
Krishna and Rao [32], 2023	Routine ROP screening cohort or acceptable image quality	—	—	—	Retrospective diagnostic accuracy study

XSL•FO
RenderX

Studies	Inclusion criteria	Exclusion criteria	Participants (N[a])	BW[b] or GA[c]	Study design
Mutua et al [33], 2023	Routine ROP screening cohort or acceptable image quality	—	—	—	Retrospective diagnostic accuracy study
Ramachandran [34], 2023	GA limit or BW restriction or routine ROP screening cohort or anatomical visualization	—	—	—	Retrospective diagnostic accuracy study
Rao et al [35], 2023	Routine ROP screening cohort or acceptable image quality or anatomical visualization	Insufficient image quality	5944	—	Retrospective diagnostic accuracy study
Sankari et al [36], 2023	Routine ROP screening cohort	Structural ocular anomalies	400	Mean BW: 1.56 (SD 0.32) kg; mean GA: 31.5 (SD 1.3) weeks	Retrospective diagnostic accuracy study
Shen et al [37], 2023	GA limit or BW restriction or routine ROP screening cohort or acceptable image quality or anatomical visualization	Insufficient image quality or discrepant annotation results	355	Mean BW: —; mean GA: 33 weeks	Retrospective diagnostic accuracy study
Subramaniam et al [38], 2023	Routine ROP screening cohort or anatomical visualization	Insufficient image quality	—	—	Retrospective diagnostic accuracy study
Young et al [39], 2023	Routine ROP screening cohort or anatomical visualization	Prior ROP treatment	156	Mean BW: 1685 (SD 419) g; mean GA: 33.0 (SD 3.0) weeks	Retrospective diagnostic accuracy study
Almeida et al [40], 2024	Routine ROP screening cohort or acceptable image quality or anatomical visualization	—	20	—	Retrospective diagnostic accuracy study
Liu et al [41], 2024	Routine ROP screening cohort or acceptable image quality	Insufficient image quality or discrepant annotation results	1089	—	Retrospective diagnostic accuracy study
Feng et al [42], 2024	Routine ROP screening cohort or acceptable image quality	Insufficient image quality	535	Mean BW: 1379.95 (SD 412.95) g (training), 1381.77 (SD 451.75) g (test); mean GA: 29.86 (SD 2.47) weeks (training), 29.29 (SD 2.94) weeks (test)	Retrospective diagnostic accuracy study
Selvamarilakshmi et al [43], 2024	Routine ROP screening cohort or acceptable image quality	—	—	—	—
Cai et al [44], 2025	GA limit or BW restriction or routine ROP screening cohort or acceptable image quality	Insufficient image quality	416	—	Retrospective diagnostic accuracy study
Chen et al [45], 2025	Routine ROP screening cohort or anatomical visualization	Stage 4-5 ROP excluded	—	—	Retrospective diagnostic accuracy study
Deepthi et al [46], 2025	Routine ROP screening cohort	—	188	—	—
Engin et al [47], 2025	GA limit or BW restriction or routine ROP screening cohort or PMA window	Insufficient image quality	355	Mean BW: 1453.19 g; mean GA: 30.11 (SD 3.8) weeks	Retrospective diagnostic accuracy study
Kaushik J et al [48], 2025	Routine ROP screening cohort or acceptable image quality or anatomical visualization	Insufficient image quality	—	—	—

XSL•FO
RenderX

Studies	Inclusion criteria	Exclusion criteria	Participants (N[a])	BW[b] or GA[c]	Study design
Kariakin et al [49], 2025	Routine ROP screening cohort or acceptable image quality	Insufficient image quality	—	—	—
Krishna and Reka [50], 2025	Routine ROP screening cohort or acceptable image quality or anatomical visualization	Insufficient image quality	—	—	—
M et al [79], 2025	Routine ROP screening cohort or anatomical visualization	—	—	—	—
Mahrous et al [51], 2025	Routine ROP screening cohort or anatomical visualization	Discrepant annotation results	—	—	—
Mowla et al [52], 2025	Routine ROP screening cohort	—	483	—	Retrospective diagnostic accuracy study
Ortiz et al [53], 2025	GA limit or BW restriction or routine ROP screening cohort or acceptable image quality or anatomical visualization	Insufficient image quality	512	Median BW: 1610 (range: 580-2800) g; median GA: 32 (range: 25-36) weeks	—
Rahim et al [54], 2025	Routine ROP screening cohort or acceptable image quality	Stage 4-5 ROP excluded	—	—	Retrospective diagnostic accuracy study
Varysova et al [55], 2025	GA limit or BW restriction or routine ROP screening cohort or anatomical visualization	Insufficient image quality	80	Mean BW: 2092.3 (range: 540-4040) g; mean GA: 33.8 (range: 24-41) weeks	Retrospective diagnostic accuracy study
Wang et al [56], 2025	Routine ROP screening cohort or anatomical visualization	—	—	—	—
Kaushik J et al [57], 2026	Routine ROP screening cohort or acceptable image quality or anatomical visualization	Insufficient image quality	—	—	—
Alam et al [58], 2025	Routine ROP screening cohort or acceptable image quality	Insufficient image quality	188	—	—
De et al [59], 2025	Routine ROP screening cohort or anatomical visualization	—	—	—	Retrospective diagnostic accuracy study
He et al [60], 2025	BW restriction or routine ROP screening cohort or PMA window	—	771	Mean BW: 852.5 (SD 208.2) g; mean GA: 26.7 (SD 1.9) weeks	Retrospective diagnostic accuracy study
Yurdakul et al [61], 2025	Routine ROP screening cohort or acceptable image quality or anatomical visualization	—	483	—	Retrospective diagnostic accuracy study
Shoeibi et al [62], 2026	Routine ROP screening cohort or anatomical visualization	Prior ROP treatment	141	Mean BW: 1197.79 (SD 494.88) g; mean GA: 28.71 (SD 3.05) weeks	Retrospective diagnostic accuracy study
Wang et al [63], 2018	Routine ROP screening cohort or anatomical visualization	Discrepant annotation results	1273	—	Retrospective diagnostic accuracy study
Ding et al [64], 2020	Routine ROP screening cohort	Insufficient image quality	—	—	Retrospective diagnostic accuracy study

XSL•FO
RenderX

Studies	Inclusion criteria	Exclusion criteria	Participants (N[a])	BW[b] or GA[c]	Study design
Coyner et al [65], 2021	Routine ROP screening cohort or acceptable image quality	Insufficient image quality	952	Mean BW: ≈949.4 g; mean GA: ≈26.9 weeks	Retrospective diagnostic accuracy study
Ramachandran et al [66], 2021	Routine ROP screening cohort or acceptable image quality	Discrepant annotation results	150	Mean BW: no-plus 1350 (SD 240) g; plus 1280 (SD 226) g; mean GA: no-plus 32.4 (SD 1.1) weeks; plus 30.9 (SD 1.8) weeks	Retrospective diagnostic accuracy study
Agrawal et al [67], 2022	GA limit or BW restriction or routine ROP screening cohort or anatomical visualization	—	—	—	Retrospective diagnostic accuracy study
Zhang et al [68], 2022	Routine ROP screening cohort or acceptable image quality	Insufficient image quality or discrepant annotation results	560	—	Retrospective diagnostic accuracy study
Redd et al [69], 2019	Routine ROP screening cohort or acceptable image quality	Insufficient image quality	870	Mean BW: 901 (SD 304) g; mean GA: 27 (SD 2) weeks	Retrospective diagnostic accuracy study
Legocki et al [70], 2023	GA limit or BW restriction or routine ROP screening cohort or acceptable image quality or anatomical visualization	—	71	Mean BW: 995.6 (SD 292.0) g; mean GA: 28.2 (SD 2.8) weeks	Retrospective diagnostic accuracy study
Derks et al [78], 2025	Routine ROP screening cohort or acceptable image quality or anatomical visualization	Insufficient image quality	66	Mean BW: 963.5 (SD 307.5) g; mean GA: 27.4 (SD 2.4) weeks	Retrospective diagnostic accuracy study
Agrawal et al [71], 2025	Routine ROP screening cohort or anatomical visualization	—	—	—	Retrospective diagnostic accuracy study
Salih et al [72], 2024	Routine ROP screening cohort or acceptable image quality	Insufficient image quality	—	—	Retrospective diagnostic accuracy study
Yang et al [73], 2025	GA limit or BW restriction or routine ROP screening cohort or PMA window	Insufficient image quality	586	Mean BW: 1.80 (SD 0.46) kg; mean GA: 31.51 (SD 1.82) weeks	Retrospective diagnostic accuracy study
Vahidmoghadam et al [74], 2026	Routine ROP screening cohort or acceptable image quality	—	188	—	Retrospective diagnostic accuracy study
Zhang et al [75], 2026	GA limit or BW restriction or routine ROP screening cohort	Insufficient image quality or prior ROP treatment	120	—	Retrospective diagnostic accuracy study
Zhao et al [76], 2026	Routine ROP screening cohort or acceptable image quality	Insufficient image quality	—	Mean BW: normal 3464.36 (SD 514.77) g; ROP 1699.81 (SD 635.21) g (ZOC); mean GA: normal 39.17 (SD 1.99) weeks; ROP 32.98 (SD 2.55) weeks (ZOC)	Retrospective diagnostic accuracy study
Huang et al [77], 2021	GA limit or BW restriction or routine ROP screening cohort or acceptable image quality or anatomical visualization	Insufficient image quality or insufficient retinal information	61	Mean BW: 1011.4 (SD 194.5)/960.8 (SD 199.5)/837.5 (SD 295.4) g (no ROP/stage1/stage2); mean GA: 28.3 (SD 2.0)/27.5 (SD1.8)/26.0 (SD 1.8) weeks	Retrospective diagnostic accuracy study

[a]N: number of infants.

[b]BW: birth weight.

[c]GA: gestational age.

[d]ROP: retinopathy of prematurity.

[e]Not available.

[f]PMA: postmenstrual age.

XSL•FO
RenderX

Two investigators (HB and YZ) independently extracted data using standardized forms. The extracted information included the first author, publication year, country or region, study setting and design, diagnostic task, model type, number of images, reference standard, validation method, and whether external validation or a comparison with ophthalmologists was reported. Sensitivity and specificity were recorded, and true positive (TP), false positive (FP), true negative (TN), and false negative (FN) counts were extracted from contingency tables or reconstructed from sufficient reported data. The extracted data were cross-checked, and discrepancies were resolved through discussion with a third investigator (BY). When a study reported multiple models, only the contingency tables for the primary model evaluated on the test dataset in the meta-analysis were included. If multiple contingency tables were available for a given model, they were assumed to be independent of each other.

Outcomes

The primary outcomes for image-based AI models were task-specific pooled sensitivity and specificity, together with the area under the summary receiver operating characteristic curve (SROC AUC). Positive and negative likelihood ratios (PLR and NLR), diagnostic odds ratio (DOR), and positive and negative predictive values (PPV and NPV) were also assessed. An overall estimate combining different diagnostic tasks and its corresponding figure are presented in Multimedia Appendix 1 as an exploratory analysis. Model and dataset characteristics are summarized in Table 2 [9,19-79]; reference standards and validation methods are presented in Table 3 [9,19-79].

XSL•FO
RenderX

Table 2. Summary of algorithms, model validation, and data sources (n=number of images).

	Algorithm		Model training and validation				Data source		
Studies	Model architecture	Method category	Training dataset	Internal validation dataset	External validation dataset	Test dataset	Source of data	Data range	Comparison between AI and clinicians[a]
Brown et al [9], 2018	U-Net and Inception version 1	DL[b]	4409	1102	85	100 (plus)	i-ROP[c] Research Consortium, comprising 8 clinical centers in the United States and Mexico	July 2011-December 2016	Yes
Hu et al [19], 2019	Inception-V2	DL	2534	400	N/A[d]	400	Chengdu Women and Children's Central Hospital	2014-2017	No
Tan et al [20], 2019	InceptionV3	DL	5579	1395	116	116	ART-ROP database, Auckland, New Zealand (4 NICUs[e]); independent clinical dataset, Hong Kong, China (External Validation)	2006-2015; N/A	No
Wang and Chen [21], 2019	InceptionV3	DL	8507	1228	N/A	1265	Sichuan Provincial People's Hospital and Chengdu Women & Children's Central Hospital, Chengdu, China	2014-2017	No
Huang et al [22], 2020	VGG19	DL	18808	N/A	N/A	N/A	Osaka Women's and Children's Hospital, Japan	Before July 2019 (start date not reported)	No
Zhang et al [23], 2019	VGG-16	DL	17801	N/A	N/A	1742	Telemed-R ROP screening system, led by Shenzhen Eye Hospital	2013-2018	No
Wang et al [24], 2021	J-PROP	DL	39029	5140	N/A	8080	Four clinical centers in Guangdong Province, China	November 4, 2010-November 14, 2019	Yes
Cole et al [25], 2022	i-ROP	DL	N/A	N/A	391; 467	N/A	Four ROP[f] screening hospitals, Kathmandu, Nepal; National Center for Maternal and Child Health, Ulaanbaatar, Mongolia	October 2016-August 2018; December 2015-January 2017	Yes
Coyner et al [26], 2022	i-ROP DL	ML[g]	1875	476	762; 330; 367	N/A	ROPE-SOS, Aravind Eye Hospitals, India; Subsequent temporal cohort from the ROPE-SOS[h] program, India; Four ROP screening hospitals in Kathmandu, Nepal; National Center for Maternal and Child Health, Ulaanbaatar, Mongolia (external validation)	February 1, 2019-December 31, 2020; January 1-June 30, 2021; October 1, 2016-August 31, 2018; December 1, 2015-January 31, 2017	Yes
Li and Liu [27], 2022	DenseNet	DL	14626	N/A	N/A	3680	Jiaxing Maternal and Child Health Hospital, Zhejiang	January 2018-December 2020	No
Luo et al [28], 2022	ResNet101	DL	750	N/A	N/A	150	Guangdong Maternal and Child Health Hospital	N/A	Yes

XSL•FO
RenderX

	Algorithm		Model training and validation				Data source		
Studies	Model architecture	Method category	Training dataset	Internal validation dataset	External validation dataset	Test dataset	Source of data	Data range	Comparison between AI and clinicians[a]
Wu et al [29], 2022	OC-Net	DL	7033	N/A	90	763	Zhujiang Hospital of Southern Medical University, Guangzhou; Second Nanning People's Hospital, Nanning (external validation)	June 3, 2017-August 28, 2019; N/A	No
Attallah [30], 2023	GabROP-SVM	ML	14,241	3560	N/A	N/A	ROP Collaboration Group (RCG[i]), China	N/A	No
Brindha et al [31], 2023	ResNet-50 + SVM	ML	27,226	3890	N/A	7779	Eye Centre, Renmin Hospital of Wuhan University, China	N/A	No
Krishna and Rao [32], 2023	DeepLabv3+	DL	15,634	N/A	N/A	3908	RCG, China	N/A	No
Mutua et al [33], 2023	Enhanced U-Net	DL	1720	215	N/A	215	Kaggle retinal image dataset	N/A	No
Ramachandran [34], 2023	U-Net + CNN[j]	DL	40,000	N/A	N/A	289	Narayana Nethralaya Eye Hospital, Bangalore, India	N/A	No
Rao et al [35], 2023	Medios AI-ROP	DL	25,982	4006	N/A	7489	KIDROP[k] tele-ROP screening program, India	2011-2022	No
Sankari et al [36], 2023	QSVM[l]	ML	3200	1000	N/A	800	SRM Hospital and Aravind Hospital	June 2019-March 2022	No
Shen et al [37], 2023	MN-ROP	DL	1534	383	N/A	N/A	Zhujiang Hospital of Southern Medical University	June 2017-November 2020	No
Subramaniam et al [38], 2023	GoogLeNet (Inception-v1)	DL	600	39	N/A	50	SP-ROP project, Argentina and Colombia	N/A	No
Young et al [39], 2023	ResNet18	DL	1065	152	N/A	305	Aravind Eye Hospital, Coimbatore	January 2021-April 2022	Yes
Almeida et al [40], 2024	DenseNet121	DL	1898	211	N/A	N/A	N/A	N/A	Yes
Liu et al [41], 2024	IRIDS[m]	DL	6158	N/A	N/A	1539	Three maternal and child health hospitals in Guangdong	April 2004-January 2021	Yes
Feng et al [42], 2024	ResNet50	DL	4796	N/A	N/A	386	Xinhua Hospital, Shanghai Jiao Tong University	N/A	Yes
Selvamarilakshmi et al [43], 2024	Random forest	ML	N/A	N/A	N/A	700	N/A	N/A	No
Cai et al [44], 2025	Swin_GDC	DL	1073	230	N/A	231	Zhujiang Hospital of Southern Medical University, Guangzhou, Guangdong Province, China	June 2017-November 2020	No
Chen et al [45], 2025	SegClass-Net	DL	1553	385	N/A	121	Guangzhou Women and Children's Medical Center, Guangzhou, Guangdong Province, China	2013-2019	No
Deepthi et al [46], 2025	TUCLM[n]	DL	1909	N/A	N/A	477	University Hospital Ostrava, Czech Republic	N/A	No

XSL•FO
RenderX

	Algorithm		Model training and validation				Data source		
Studies	Model architecture	Method category	Training dataset	Internal validation dataset	External validation dataset	Test dataset	Source of data	Data range	Comparison between AI and clinicians[a]
Engin et al [47], 2025	XGBoost[o]	ML	N/A	54	N/A	N/A	Dokuz Eylul University Hospital	N/A	No
Kaushik J et al [48], 2025	ConvNeXT	DL	720	N/A	N/A	180 (plus)	Narayana Nethralaya Eye Hospital	N/A	No
Kariakin et al [49], 2025	ViT	DL	3570	714	N/A	892	Multiprofile Clinical Medical Center "Bonum," Sverdlovsk Region, Russia	N/A	No
Krishna and Reka [50], 2025	Soft voting or mixed voting	DL	N/A	N/A	N/A	367; 284	Aravind Eye Hospital, Chennai, India	N/A	No
M et al [79], 2025	Vision transformer	DL	N/A	N/A	N/A	213	Retinopathy of Prematurity Database (ROPDB)	N/A	No
Mahrous et al [51], 2025	ResNet50	DL	2425	270	N/A	675	Kaggle retinal image dataset of infants and ROP	N/A	No
Mowla et al [52], 2025	LightEyeNet	DL	3224	403	N/A	403	Shenzhen Eye Hospital	2004-2023	No
Ortiz et al [53], 2025	ResNet18 ROP	DL	1664	272	N/A	291	Participating hospitals in Mexico and Argentina	May 12, 2020-October 31, 2023	No
Rahim et al [54], 2025	DPFRr-CLAHE + InceptionResNetV2	DL	1422	356	N/A	N/A	Clinical retinal images collected in collaboration with the University of Calgary and Dr Anna Ells	N/A	Yes
Varysova et al [55], 2025	ResNet18	DL	160	40	N/A	N/A	University Hospital Ostrava	2019-2020	Yes
Wang et al [56], 2025	ResNet-34	DL	736	N/A	N/A	184	Department of Ophthalmology, Jilin Provincial Maternal and Child Health Hospital	N/A	No
Kaushik J et al [57], 2026	ResNet50	DL	1600	N/A	N/A	400 (ROP)	Narayana Nethralaya Eye Hospital, Bengaluru, Karnataka, India	N/A	No
Alam et al [58], 2025	CNN	DL	9874	1234	N/A	1234	Kaggle	N/A	No
De et al [59], 2025	QROP[p]	ML	272	34	N/A	34	H.V. Desai Eye Hospital, India	2023	No
He et al [60], 2025	DNN[q]	DL	N/A	N/A	N/A	771	e-ROP[r] study, 13 North American centers	N/A	No
Yurdakul et al [61], 2025	ROPGCViT	DL	877	109	N/A	113	IROPS retinal fundus image dataset	2004-2023	No
Shoeibi et al [62], 2025	MobileNet with CLAHE[s]	DL	1360	N/A	N/A	340	ROP Clinic Telemedicine Consultation System, Khatam-Al-Anbia Eye Hospital	2021-2024	No
Wang et al [63], 2018	Id-Net and Gr-Net	DL	13526; 4139	2361	944; 106	4908	Chengdu Women & Children's Central Hospital, China	March 3-December 28, 2017	No
Ding et al [64], 2020	Mask R-CNN + Inception v3 hybrid	DL	1657	274	N/A	828	Hospital-based retinal image dataset; hospital unspecified	N/A	Yes

XSL•FO
RenderX

	Algorithm		Model training and validation				Data source		
Studies	Model architecture	Method category	Training dataset	Internal validation dataset	External validation dataset	Test dataset	Source of data	Data range	Comparison between AI and clinicians[a]
Coyner et al [65], 2021	i-ROP	ML	376	N/A	132	444	Multicenter imaging and Informatics in Retinopathy of Prematurity (i-ROP) cohort	September 2015-June 2018	Yes
Ramachandran et al [66], 2021	YOLO[t]-based fully CNN	DL	384	77	N/A	161	Karnataka Internet Assisted Diagnosis of Retinopathy of Prematurity (KIDROP)	N/A	No
Agrawal et al [67], 2022	AG[u] U-Net	DL	648	55	N/A	200 (ROP)	PBMA's[v] H.V. Desai Eye Hospital, Pune	N/A	No
Zhang et al [68], 2022	Combined SE-HBP50 network	DL	6867	1709	N/A	3654	Shenzhen Eye Hospital, China	2009-2018	No
Redd et al [69], 2019	GoogLeNet	DL	4409	4104	N/A	100 (TR-ROP[w])	i-ROP Research Consortium, comprising 8 clinical centers in the United States and Mexico	July 2011-December 2016	Yes
Legocki et al [70], 2023	XGBoost	ML	167	N/A	N/A	N/A	University of Washington and Seattle Children's Hospital NICUs	July 2015-February 2018	Yes
Derks et al [78], 2025	i-ROP DL algorithm	DL	N/A	N/A	213	1223	Erasmus University Medical Center, Rotterdam	N/A	No
Agrawal et al [71], 2025	MultiC-NN_DNN / MultiC-NN_LSTM	DL	617	88	N/A	177 (Stage)	H.V. Desai Eye Hospital, Pune, India	N/A	No
Salih et al [72], 2024	ResNet50, DenseNet161	DL	1942	278	N/A	556	Al-Amal Eye Center Private Clinic, Baghdad, Iraq	2015-2020	No
Yang et al [73], 2025	Random forest	ML	N/A	175	N/A	N/A	First Affiliated Hospital of Guangxi Medical University	January 2019-January 2024	No
Vahidmoghadam et al [74], 2026	U-Net++	DL	4803 (plus); 703 (Stage)	960 (plus); 140 (stage)	N/A	241 (plus); 37 (stage)	University Hospital Ostrava, Czech Republic (public dataset)	N/A	No
Zhang et al [75], 2026	ResNet50	DL	2281	254	100	100	Iran: FARFUM-RoP; China: LAN-RoP; India: HVDROPDB	FARFUM-RoP: N/A; LAN-RoP: January 2019-Mar 2025; HVDROPDB: 2020-2023	No
Zhao et al [76], 2026	DeepAdapter	DL	13,544	5805	19,153; 954	19153; 954	China: ZOC[x], LZH[y], LYH[z] (development); GDH[aa] (external); India and Czech Republic (public external)	N/A	Yes

XSL•FO
RenderX

Studies	Algorithm		Model training and validation				Data source		
	Model architecture	Method category	Training dataset	Internal validation dataset	External validation dataset	Test dataset	Source of data	Data range	Comparison between AI and clinicians[a]
Huang et al [77], 2021	CNN	DL	10,235	1137	N/A	244	Neonatal Intensive Care Unit of Chang Gung Memorial Hospital, Linkou, Taiwan	December 17, 2013-May 24, 2019	No

[a]"Yes" under comparison between AI and clinicians indicates that the study reported clinician assessment results. It does not necessarily mean that clinicians and AI were evaluated on the same dataset or that the comparison data were extractable for meta-analysis.

[b]DL: deep learning.

[c]i-ROP: Informatics in Retinopathy of Prematurity.

[d]N/A: not applicable.

[e]NICU: neonatal intensive care unit.

[f]ROP: retinopathy of prematurity.

[g]ML: machine learning.

[h]ROPE-SOS: Retinopathy of Prematurity Eradication–Save Our Sight.

[i]RCG: ROP Collaboration Group.

[j]CNN: convolutional neural network.

[k]KIDROP: Karnataka Internet Assisted Diagnosis of Retinopathy of Prematurity.

[l]QSVM: quantum support vector machine.

[m]IRIDS: Infant Retinal Intelligent Diagnosis System.

[n]TUCLM: transformer-based unsupervised curriculum learning model.

[o]XGBoost: extreme gradient boosting.

[p]QROP: quantum retinopathy of prematurity system.

[q]DNN: deep neural network.

[r]e-ROP: Telemedicine Approaches to Evaluating Acute-Phase Retinopathy of Prematurity.

[s]CLAHE: contrast-limited adaptive histogram equalization.

[t]YOLO: you only look once.

[u]AG: attention gate.

[v]PBMA: Poona Blind Men's Association.

[w]TR-ROP: treatment-requiring retinopathy of prematurity.

[x]ZOC: Zhongshan Ophthalmic Center.

[y]LZH: Liuzhou Maternity and Child Healthcare Hospital.

[z]LYH: Linyi Maternity and Child Healthcare Hospital.

[aa]GDH: Guangdong Women and Children Hospital.

XSL•FO
RenderX

Table 3. Reference standards and validation methods of the included studies[a].

Studies	Reference standard	Type of internal validation	External validation	Transfer learning applied
Brown et al [9], 2018	Expert consensus	Hold-out validation or 5-fold cross-validation	Yes	Yes
Hu et al [19], 2019	Expert consensus	Hold-out validation	No	Yes
Tan et al [20], 2019	Clinical diagnosis	Hold-out validation	Yes	No
Wang and Chen [21], 2019	Expert consensus	Hold-out validation	No	Yes
Huang et al [22], 2020	Expert consensus	Hold-out validation or 5-fold cross-validation	No	Yes
Zhang et al [23], 2019	Expert consensus	Hold-out validation	No	Yes
Wang et al [24], 2021	Expert consensus	Hold-out validation	No	Yes
Cole et al [25], 2022	Expert consensus	Hold-out validation	Yes	No
Coyner et al [26], 2022	Expert consensus	Hold-out validation	Yes	Yes
Li and Liu [27], 2022	Expert consensus	Hold-out validation or 5-fold cross-validation	No	Yes
Luo et al [28], 2022	Expert consensus	Hold-out validation	No	No
Wu et al [29], 2022	Expert consensus	5-fold cross-validation	Yes	No
Attallah [30], 2023	Expert consensus	5-fold cross-validation	No	Yes
Brindha et al [31], 2023	N/A[b]	Hold-out validation	No	No
Krishna and Rao [32], 2023	Expert consensus	Hold-out validation	No	No
Mutua et al [33], 2023	Expert consensus	Hold-out validation	No	Yes
Ramachandran [34], 2023	Expert consensus	Hold-out validation	No	No
Rao et al [35], 2023	Expert consensus	Hold-out validation	No	Yes
Sankari et al [36], 2023	Expert consensus	5-fold cross-validation	No	No
Shen et al [37], 2023	Expert consensus	5-fold cross-validation	No	Yes
Subramaniam et al [38], 2023	Expert consensus	Hold-out validation or 5-fold cross-validation	No	Yes
Young et al [39], 2023	Expert consensus	Hold-out validation	No	No
Almeida et al [40], 2024	Expert consensus	10-fold cross-validation	No	Yes
Liu et al [41], 2024	Expert consensus	5-fold cross-validation	No	No
Feng et al [42], 2024	Expert consensus	Hold-out validation	No	Yes
Selvamarilakshmi D. et al [43], 2024	Expert consensus	K-fold cross-validation	No	No
Cai et al [44], 2025	Labels from hospital or public dataset	5-fold cross-validation	No	Yes
Chen et al [45], 2025	Expert consensus	Hold-out validation or 5-fold cross-validation	No	Yes
Deepthi et al [46], 2025	Labels from hospital or public dataset	Hold-out validation	No	Yes
Engin et al [47], 2025	Expert consensus	Hold-out validation	No	No
Kaushik J et al [48], 2025	Expert consensus	5-fold cross-validation	No	Yes
Kariakin et al [49], 2025	Labels from hospital or public dataset	Hold-out validation or 5-fold cross-validation	No	Yes
Krishna and Reka [50], 2025	Labels from hospital or public dataset	5-fold cross-validation	No	Yes
M et al [79], 2025	Expert consensus	Hold-out validation	No	Yes

XSL•FO
RenderX

Studies	Reference standard	Type of internal validation	External validation	Transfer learning applied
Mahrous et al [51], 2025	Labels from hospital or public dataset	Hold-out validation	No	No
Mowla et al [52], 2025	Labels from hospital or public dataset	Hold-out validation	No	Yes
Ortiz et al [53], 2025	Expert consensus	Hold-out validation	No	No
Rahim et al [54], 2025	Labels from hospital or public dataset	Hold-out validation or 5-fold cross-validation	No	Yes
Varysova et al [55], 2025	Expert consensus	Hold-out validation	No	No
Wang et al [56], 2025	Expert consensus	Hold-out validation	No	Yes
Kaushik J et al [57], 2026	Labels from hospital or public dataset	Hold-out validation	No	Yes
Alam et al [58], 2025	Expert consensus	Hold-out validation	No	No
De et al [59], 2025	Expert consensus	Hold-out validation	No	No
He et al [60], 2025	Expert consensus	10-fold cross-validation	No	No
Yurdakul et al [61], 2025	Labels from hospital or public dataset	Hold-out validation	No	Yes
Shoeibi et al [62], 2025	Labels from hospital or public dataset	Hold-out validation	No	Yes
Wang et al [63], 2018	Expert consensus	Hold-out validation	Yes	Yes
Ding et al [64], 2020	Expert consensus	Hold-out validation	No	Yes
Coyner et al [65], 2021	Expert consensus	5-fold cross-validation	Yes	No
Ramachandran et al [66], 2021	Expert consensus	Hold-out validation	No	Yes
Agrawal et al [67], 2022	Expert consensus	Hold-out validation	No	No
Zhang et al [68], 2022	Expert consensus	Hold-out validation	No	Yes
Redd et al [69], 2019	Expert consensus	5-fold cross-validation	No	Yes
Legocki et al [70], 2023	Labels from hospital or public dataset	Hold-out validation	No	No
Derks et al [78], 2025	Expert consensus	Hold-out validation	Yes	No
Agrawal et al [71], 2025	Expert consensus	Hold-out validation	No	Yes
Salih et al [72], 2024	Expert consensus	Hold-out validation	No	Yes
Yang et al [73], 2025	Labels from hospital or public dataset	Hold-out validation or 5-fold cross-validation	No	No
Vahidmoghadam et al [74], 2026	Expert consensus	Hold-out validation or 5-fold cross-validation	No	Yes
Zhang et al [75], 2026	Expert consensus	10-fold cross-validation	Yes	Yes
Zhao et al [76], 2026	Expert consensus	Hold-out validation	Yes	Yes
Huang et al [77], 2021	Expert consensus	5-fold cross-validation	No	No

[a]"Yes" under external validation indicates that the study conducted external validation and reported extractable data.

[b]N/A: not applicable.

Quality Assessment

Two reviewers (HB and YZ) independently assessed the included studies for risk of bias (RoB) and applicability concerns using the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) tool [80]. A third investigator (BY) resolved disagreements when necessary. Domain-specific percentage agreement and unweighted Cohen κ were calculated from the two investigators' initial judgments before adjudication, and the results are reported in Table S2 in Multimedia Appendix 1 [81]. The final QUADAS-2 judgments were recorded and visualized using Review Manager 5.4 (The Cochrane Collaboration). The certainty of the body of evidence was evaluated using the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) system [82].

XSL•FO
RenderX

We used QUADAS-2 because this review primarily synthesized 2×2 contingency-table data to evaluate the diagnostic accuracy of image-based AI for ROP against clinical reference standards. Its domains of patient selection, index test, reference standard, and flow and timing address the principal risk-of-bias and applicability concerns in the included diagnostic accuracy studies. QUADAS-AI (Quality Assessment of Diagnostic Accuracy Studies–Artificial Intelligence) remains under development, whereas PROBAST+AI (Prediction Model Risk of Bias Assessment Tool + Artificial Intelligence) focuses on the development and performance evaluation of prediction models.

Statistical Analysis

For each included study, we selected the authors' primary model and extracted or reconstructed image-level counts of TP, FP, FN, and TN from sufficiently reported data to construct 2×2 contingency tables. Sensitivity and specificity were jointly pooled using a bivariate mixed-effects model; PLRs, NLRs, and DORs were derived from the model estimates [83]. We generated an SROC curve and estimated its area under the curve (AUC). Pooled sensitivity and specificity were reported with 95% CIs, and 95% confidence and prediction regions were displayed on the SROC plot. Heterogeneity in sensitivity and specificity was assessed separately using I^2 statistics. Subgroup meta-analyses and meta-regression were performed to explore potential sources of heterogeneity. Potential publication bias was assessed using the Deeks funnel plot asymmetry test, supplemented by trim-and-fill and contour-enhanced funnel plots and the Egger test, and Fagan nomograms were used to assess clinical applicability across the included studies. Meta-regression analyses were performed using R (version 4.5.3; R Foundation) [84], whereas all other meta-analyses were conducted using Stata (version 18.0; StataCorp LLC). Statistical significance was set at $P<.05$.

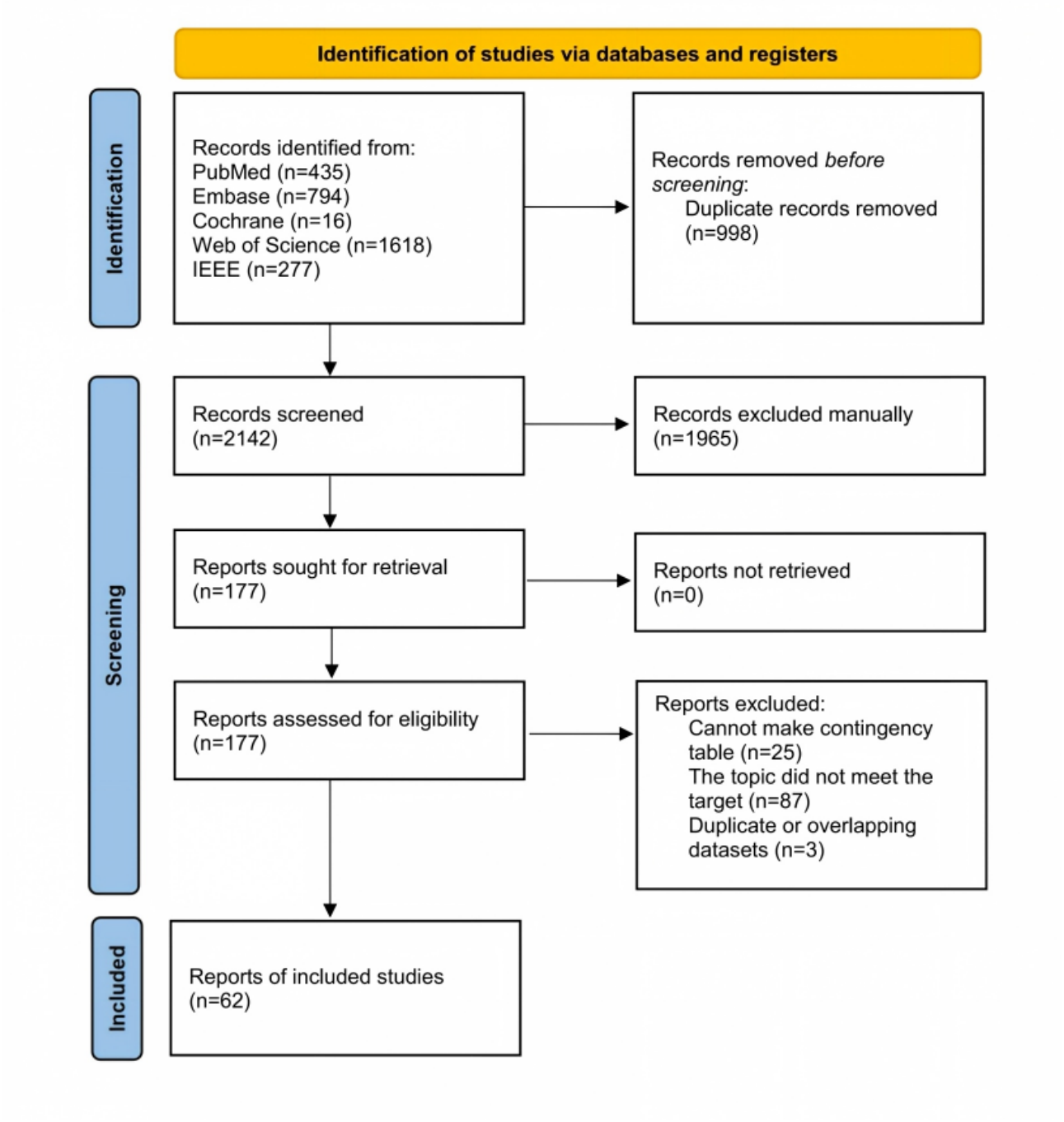
Results

Study Selection

The initial search identified 3140 records. After removal of 998 duplicates, 2142 records underwent title and abstract screening, of which 1965 were excluded because their topics or study designs did not meet the eligibility criteria. The remaining 177 reports underwent full-text assessment. Of these, 115 were excluded: 25 lacked sufficient data to construct a 2×2 contingency table; 87 were outside the study scope, including reports that did not focus on image-based AI models for ROP detection, screening, staging, or severity classification; and 3 used overlapping datasets to evaluate the same diagnostic task. For studies with overlapping datasets, we prioritized those with independent test sets and no evidence of training-test data leakage, followed by those with larger and more representative test populations. Ultimately, 62 studies were included, all of which contributed to the quantitative meta-analysis (Figure 1).

XSL•FO
RenderX

Figure 1. PRISMA-DTA (Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies) flowchart (database search, deduplication, title and abstract screening, full-text examination, and final inclusion). Sixty-two studies were ultimately included.



Study Characteristics

As shown in Tables 1-3, the 62 included studies were published between 2018 and 2026. Fifty-one were classified as retrospective diagnostic accuracy studies. All studies reported inclusion criteria, and 35 excluded images of insufficient quality. Most analyzed fundus images from routine ROP screening cohorts. Participant counts were reported in 38 studies and summed to 63,117 across studies; the remaining 24 did not clearly report participant numbers. Expert consensus was the main reference standard (n=48). DL models predominated (n=52), chiefly using convolutional neural networks, and 38 studies applied TL. Ten studies reported external validation, and 3 studies compared the diagnostic performance of AI models and experts using the same dataset.

Meta-Analysis Overview

A total of 62 studies, contributing 185 contingency tables, provided sufficient data for the diagnostic accuracy meta-analysis (Table S3 in Multimedia Appendix 1). Detailed results for the overall pooled diagnostic performance of the AI models are presented in Figures S1 and S2 in Multimedia Appendix 1.

To characterize variation in diagnostic accuracy, we conducted subgroup analyses by algorithm type, diagnostic task, and validation method. Algorithms were classified as DL or conventional ML according to their reported architectures.

XSL•FO
RenderX

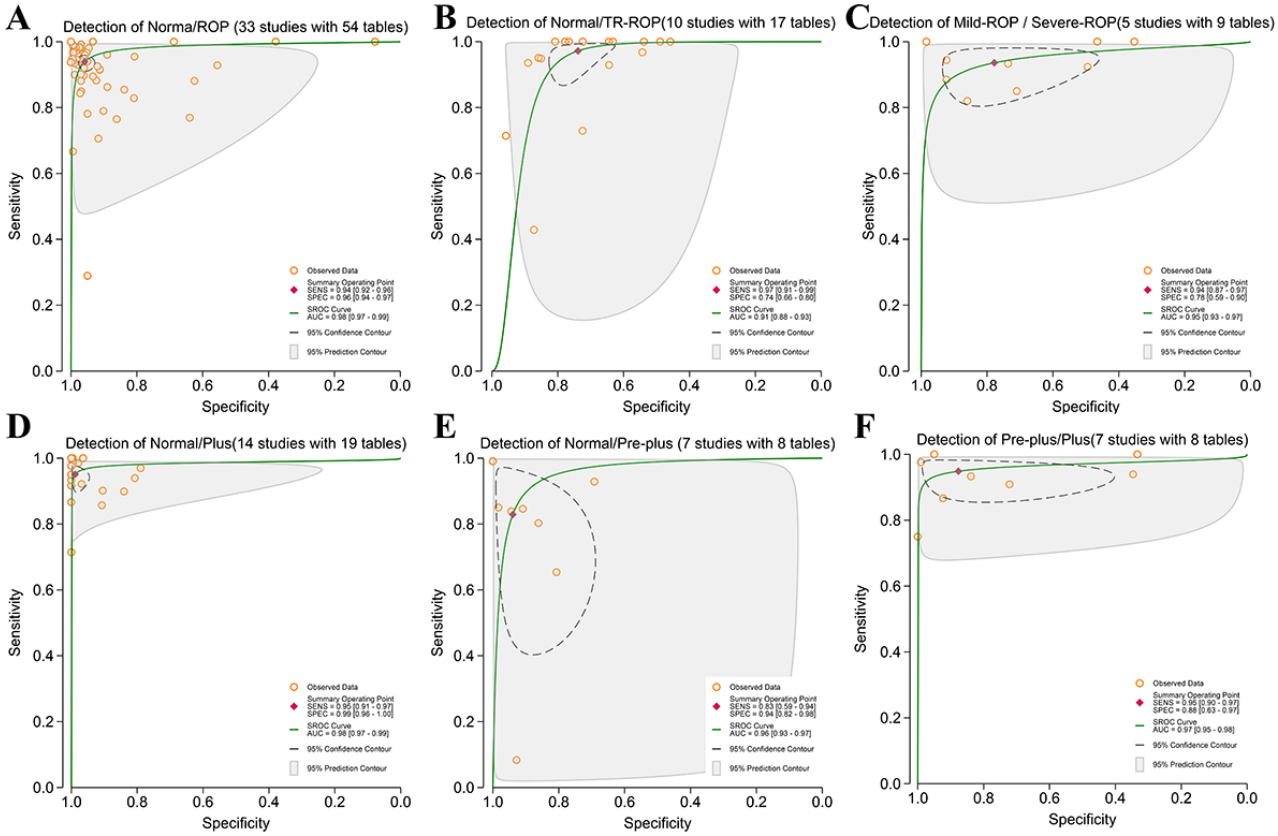
Diagnostic tasks were grouped with reference to the International Classification of Retinopathy of Prematurity, Third Edition (ICROP3), and published clinical and image-based classification frameworks [4]. Task-specific analyses covered detection of ROP (normal/ROP), treatment-requiring ROP (normal/TR-ROP), plus disease classification (normal/preplus, normal/plus, and preplus/plus), disease staging (normal/stages 1, 2, or 3 and pairwise comparisons among these stages), and severity classification (mild ROP/severe ROP) [85]. Validation methods included internal hold-out validation, cross-validation, and external validation [86]. Only test folds that were used in turn during a cross-validation procedure were classified as cross-validation; their corresponding performance estimates were also not included in the internal-validation subgroup.

Subgroup Analyses

Disease Types

The normal/ROP diagnostic task included 33 studies (54 contingency tables), with a pooled sensitivity of 0.94 (95% CI 0.92-0.96), a pooled specificity of 0.96 (95% CI 0.94-0.97), and an AUC of 0.98 (95% CI 0.97-0.99; Figure 2A). The pooled PLR, NLR, and DOR were 22.6 (95% CI 14.5-35.4), 0.06 (95% CI 0.05-0.09), and 355 (95% CI 194-649), respectively. The Deeks funnel plot demonstrated no significant publication bias (*P*=.75; Figure S3 in Multimedia Appendix 1).

Figure 2. SROC curves stratified by disease type and diagnostic task: (A) normal/ROP (33 studies, 54 contingency tables), (B) normal/TR-ROP (10 studies, 17 contingency tables), (C) mild/severe ROP (5 studies, 9 contingency tables), (D) normal/plus disease (14 studies, 19 contingency tables), (E) normal/preplus disease (7 studies, 8 contingency tables), (F) preplus/plus disease (7 studies, 8 contingency tables). Summary operating points, 95% CIs, and 95% prediction regions are displayed. AUC: area under the curve; ROP: retinopathy of prematurity; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic; TR-ROP: treatment-requiring retinopathy of prematurity.



The normal/TR-ROP diagnostic task included 10 studies (17 contingency tables), with a pooled sensitivity of 0.97 (95% CI 0.91-0.99), a pooled specificity of 0.74 (95% CI 0.66-0.80), and an AUC of 0.91 (95% CI 0.88-0.93; Figure 2B). The pooled PLR, NLR, and DOR were 3.7 (95% CI 2.9-4.8), 0.04 (95% CI 0.01-0.12), and 98 (95% CI 33-292), respectively. The Deeks funnel plot demonstrated no significant publication bias (*P*=.53; Figure S4 in Multimedia Appendix 1).

The mild ROP/severe ROP diagnostic task included 5 studies (9 contingency tables), with a pooled sensitivity of 0.94 (95% CI 0.87-0.97), a pooled specificity of 0.78 (95% CI 0.59-0.90), and an AUC of 0.95 (95% CI 0.93-0.97; Figure 2C). The pooled PLR, NLR, and DOR were 4.2 (95% CI 2.1-8.4), 0.08 (95% CI 0.04-0.17), and 51 (95% CI 17-125), respectively. The Deeks funnel plot demonstrated no significant publication bias (*P*=.52; Figure S5 in Multimedia Appendix 1).

For normal/plus disease detection, 14 studies comprising 19 contingency tables yielded a pooled sensitivity of 0.95 (95% CI 0.91-0.97), specificity of 0.99 (95% CI 0.96-1.00), and an AUC of 0.98 (95% CI 0.97-0.99; Figure 2D). The pooled PLR, NLR, and DOR were 80.0 (95% CI 25.4-252.0), 0.05 (95% CI 0.03-0.09), and 1611 (95% CI 392-6627), respectively. The Deeks funnel plot demonstrated no significant publication bias (*P*=.52; Figure S6 in Multimedia Appendix 1).

For normal/preplus disease detection, 7 studies comprising 8 contingency tables yielded a pooled sensitivity of 0.83 (95%

XSL•FO
RenderX

CI 0.59-0.94), a specificity of 0.94 (95% CI 0.82-0.98), and an AUC of 0.96 (95% CI 0.93-0.97; Figure 2E). The pooled PLR, NLR, and DOR were 13.7 (95% CI 4.1-46.2), 0.18 (95% CI 0.07-0.51), and 75 (95% CI 10-556), respectively. The Deeks funnel plot demonstrated no significant publication bias (*P*=.36; Figure S7 in Multimedia Appendix 1).

For preplus/plus disease classification, 7 studies comprising 8 contingency tables yielded a pooled sensitivity of 0.95 (95% CI 0.90-0.97), a specificity of 0.88 (95% CI 0.63-0.97), and an AUC of 0.97 (95% CI 0.95-0.98; Figure 2F). The pooled PLR, NLR, and DOR were 7.7 (95% CI 2.2-27.4), 0.06 (95% CI 0.03-0.12), and 132 (95% CI 23-763), respectively. The Deeks funnel plot demonstrated no significant publication bias (*P*=.16; Figure S8 in Multimedia Appendix 1).

The normal/stage 1 diagnostic task included 10 studies (11 contingency tables), and the pooled sensitivity, specificity, and AUC were 0.98 (95% CI 0.90-0.99), 0.98 (95% CI 0.94-1.00), and 1.00 (95% CI 0.99-1.00), respectively (Figure S9 in Multimedia Appendix 1). The pooled PLR, NLR, and DOR were 57.7 (95% CI 14.7-227.3), 0.03 (95% CI 0.01-0.10), and 2271 (95% CI 301-17,118), respectively. The Deeks funnel plot demonstrated no significant publication bias (*P*=.97; Figure S10 in Multimedia Appendix 1).

The normal/stage 2 diagnostic task included 10 studies (11 contingency tables), and the pooled sensitivity, specificity, and AUC were 0.98 (95% CI 0.90-1.00), 0.99 (95% CI 0.97-1.00), and 1.00 (95% CI 0.99-1.00), respectively (Figure S11 in Multimedia Appendix 1). The pooled PLR, NLR, and DOR were 112.6 (95% CI 28.3-448.5), 0.02 (95% CI 0.00-0.11), and 5758 (95% CI 689-48097), respectively. The Deeks funnel plot demonstrated no significant publication bias (*P*=.18; Figure S12 in Multimedia Appendix 1).

The normal/stage 3 diagnostic task included 10 studies (11 contingency tables), and the pooled sensitivity, specificity, and AUC were 0.95 (95% CI 0.85-0.99), 0.99 (95% CI 0.97-1.00), and 1.00 (95% CI 0.99-1.00), respectively (Figure S13 in Multimedia Appendix 1). The pooled PLR, NLR, and DOR were 113.2 (95% CI 30.7-416.6), 0.05 (95% CI 0.01-0.16), and 2372 (95% CI 281-20,059), respectively. The Deeks funnel plot demonstrated no significant publication bias (*P*=.26; Figure S14 in Multimedia Appendix 1).

The stage 1/stage 2 diagnostic task included 12 studies (13 contingency tables), and the pooled sensitivity, specificity, and AUC were 0.97 (95% CI 0.88-0.99), 0.96 (95% CI 0.91-0.98), and 0.99 (95% CI 0.97-0.99), respectively (Figure S15 in Multimedia Appendix 1). The pooled PLR, NLR, and DOR were 21.7 (95% CI 10.4-45.0), 0.03 (95% CI 0.01-0.13), and 689 (95% CI 105-4538), respectively. The Deeks funnel plot demonstrated no significant publication bias (*P*=.28; Figure S16 in Multimedia Appendix 1).

The stage 1/stage 3 diagnostic task included 11 studies (12 contingency tables), and the pooled sensitivity, specificity, and AUC were 0.95 (95% CI 0.83-0.99), 0.97 (95% CI 0.92-0.99), and 0.99 (95% CI 0.98-1.00), respectively (Figure S17 in Multimedia Appendix 1). The pooled PLR, NLR, and DOR were 30.3 (95% CI 11.8-77.9), 0.05 (95% CI 0.02-0.19), and 569 (95% CI 87-3724), respectively. The Deeks funnel plot demonstrated no significant publication bias (*P*=.70; Figure S18 in Multimedia Appendix 1).

The stage 2/stage 3 diagnostic task included 11 studies (12 contingency tables), and the pooled sensitivity, specificity, and AUC were 0.96 (95% CI 0.87-0.99), 0.96 (95% CI 0.90-0.99), and 0.99 (95% CI 0.98-1.00), respectively (Figure S19 in Multimedia Appendix 1). The pooled PLR, NLR, and DOR were 26.0 (95% CI 8.9-76.0), 0.04 (95% CI 0.01-0.14), and 589 (95% CI 76-4577), respectively. The Deeks funnel plot demonstrated no significant publication bias (*P*=.15; Figure S20 in Multimedia Appendix 1).

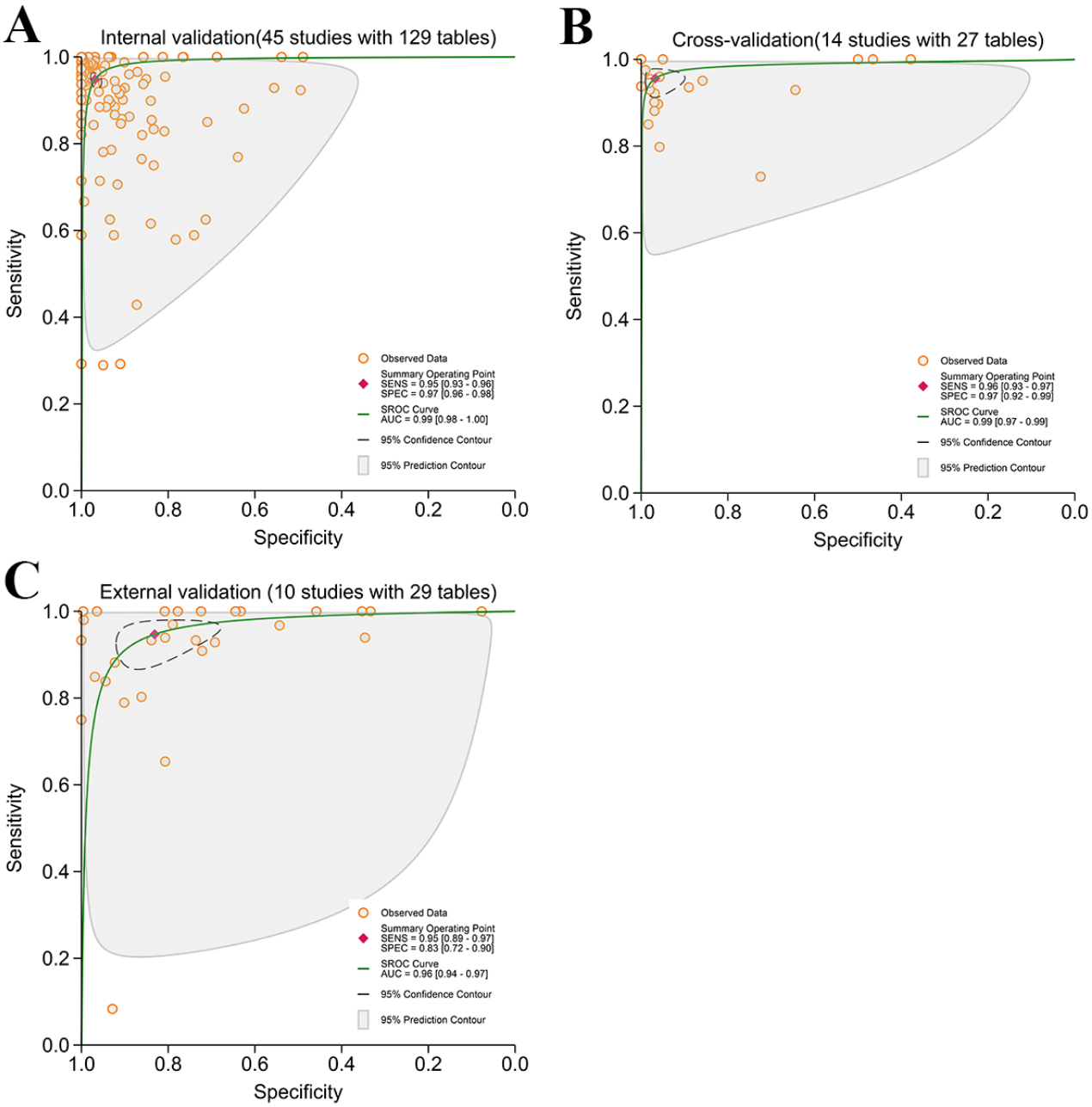
The PPV and NPV for each disease type are provided in Figures S21-S32 in Multimedia Appendix 1.

Validation Methods

The internal validation subgroup included 45 studies (129 contingency tables), with a pooled sensitivity of 0.95 (95% CI 0.93-0.96), a pooled specificity of 0.97 (95% CI 0.96-0.98), and an AUC of 0.99 (95% CI 0.98-1.00; Figure 3A). The pooled PLR, NLR, and DOR were 30.0 (95% CI 21.8-41.3), 0.05 (95% CI 0.04-0.07), and 563 (95% CI 340-932), respectively. The Deeks funnel plot revealed no significant publication bias (*P*=.11; Figure S33 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 3. SROC curves stratified by validation method: (A) internal validation (45 studies, 129 contingency tables), (B) cross-validation (14 studies, 27 contingency tables), (C) external validation (10 studies, 29 contingency tables). Summary operating points, AUCs, 95% CIs, and 95% prediction regions are displayed. AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.



The cross-validation subgroup included 14 studies (27 contingency tables), with a pooled sensitivity of 0.96 (95% CI 0.93-0.97), a pooled specificity of 0.97 (95% CI 0.92-0.99), and an AUC of 0.99 (95% CI 0.97-0.99; Figure 3B). The pooled PLR, NLR, and DOR were 29.5 (95% CI 12.2-71.4), 0.05 (95% CI 0.03-0.08), and 646 (95% CI 226-1845), respectively. The Deeks funnel plot revealed no significant publication bias (*P*=.84; Figure S34 in Multimedia Appendix 1).

The external validation subgroup included 10 studies (29 contingency tables), with a pooled sensitivity of 0.95 (95% CI 0.89-0.97), a pooled specificity of 0.83 (95% CI 0.72-0.90), and an AUC of 0.96 (95% CI 0.94-0.97; Figure 3C). The pooled PLR, NLR, and DOR were 5.6 (95% CI 3.3-9.5), 0.06 (95% CI 0.03-0.13), and 88 (95% CI 39-197), respectively. The Deeks funnel plot revealed no significant publication bias (*P*=.40; Figure S35 in Multimedia Appendix 1).

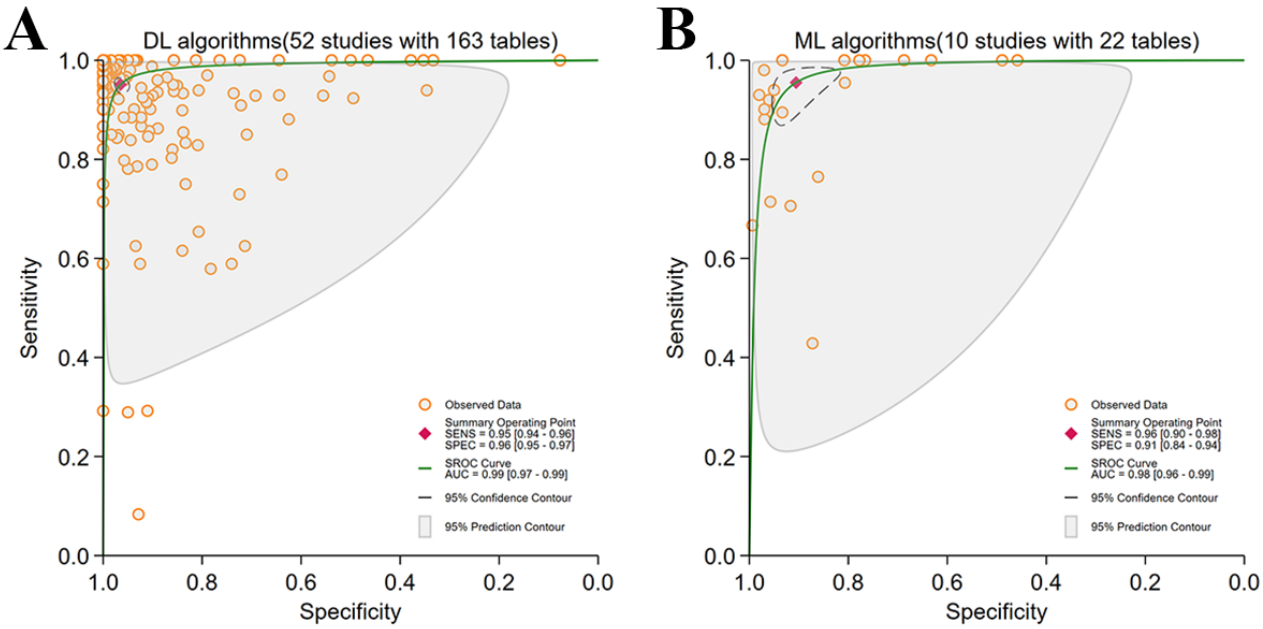
The PPV and NPV for each validation method are provided in Figures S36-S38 in Multimedia Appendix 1.

Model Types

A total of 52 studies (163 contingency tables) were included in the DL subgroup, and their pooled sensitivity, specificity, and AUC were 0.95 (95% CI 0.94-0.96), 0.96 (95% CI 0.95-0.97), and 0.99 (95% CI 0.97-0.99), respectively (Figure 4A). The pooled PLR, NLR, and DOR were 26.9 (95% CI 19.3-37.5), 0.05 (95% CI 0.04-0.07), and 532 (95% CI 330-858), respectively. The Deeks funnel plot demonstrated no significant publication bias (*P*=.47; Figure S39 in Multimedia Appendix 1).

XSL•FO
RenderX

Figure 4. SROC curves stratified by model type: (A) DL (52 studies, 163 contingency tables), (B) ML (10 studies, 22 contingency tables). Summary operating points, AUCs, 95% CIs, and 95% prediction regions are displayed. AUC: area under the curve; DL: deep learning; ML: machine learning; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.



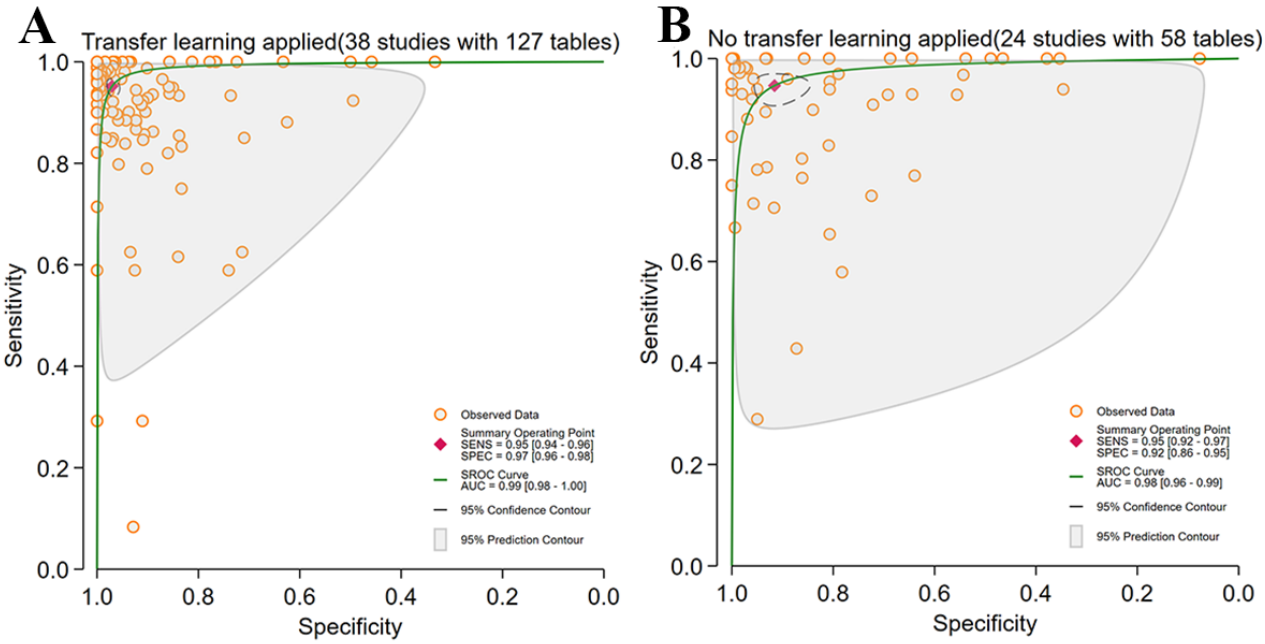
A total of 10 studies (22 contingency tables) were included in the ML subgroup, and their pooled sensitivity, specificity, and AUC were 0.96 (95% CI 0.90-0.98), 0.91 (95% CI 0.84-0.94), and 0.98 (95% CI 0.96-0.99), respectively (Figure 4B). The pooled PLR, NLR, and DOR were 10.1 (95% CI 6.1-16.6), 0.05 (95% CI 0.02-0.11), and 204 (95% CI 97-428), respectively. The Deeks funnel plot demonstrated no significant publication bias (P=.42; Figure S40 in Multimedia Appendix 1).

The PPV and NPV for each model type are provided in Figures S41 and S42 in Multimedia Appendix 1.

TL

The TL-applied subgroup included 38 studies (127 contingency tables), and their pooled sensitivity, specificity, and AUC were 0.95 (95% CI 0.94-0.96), 0.97 (95% CI 0.96-0.98), and 0.99 (95% CI 0.98-1.00), respectively (Figure 5A). The pooled PLR, NLR, and DOR were 32.3 (95% CI 23.0-45.2), 0.05 (95% CI 0.04-0.07), and 649 (95% CI 391-1078), respectively. Deeks funnel plot demonstrated significant publication bias (P=.04; Figure S43 in Multimedia Appendix 1).

Figure 5. SROC curves for groups with transfer learning (TL) applied or not: (A) TL applied (38 studies, 127 contingency tables), (B) TL not applied (24 studies, 58 contingency tables). Summary operating points, AUCs, 95% CIs, and 95% prediction regions are displayed. AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.



XSL•FO
RenderX

The no TL–applied subgroup included 24 studies (58 contingency tables), and their pooled sensitivity, specificity, and AUC were 0.95 (95% CI 0.92-0.97), 0.92 (95% CI 0.86-0.95), and 0.98 (95% CI 0.96-0.99), respectively (Figure 5B). The pooled PLR, NLR, and DOR were 11.2 (95% CI 6.9-18.4), 0.06 (95% CI 0.04-0.09), and 193 (95% CI 96-389), respectively. The Deeks funnel plot demonstrated significant publication bias (P<.001; Figure S44 in Multimedia Appendix 1).

The PPV and NPV for "TL-applied" and "no TL–applied" are provided in Figures S45 and S46 in Multimedia Appendix 1.

AI Versus Human Experts

Three studies compared AI models with human experts on the same dataset, with 4 contingency tables for AI and 14 contingency tables for human experts (Figure S47 in Multimedia Appendix 1).

The pooled sensitivity, specificity, and AUC were 0.91 (95% CI 0.83-0.95), 0.97 (95% CI 0.87-0.99), and 0.96 (95% CI 0.94-0.98) for AI models, and 0.96 (95% CI 0.91-0.99), 0.99 (95% CI 0.98-1.00), and 1.00 (95% CI 0.99-1.00) for human experts. The pooled PLR, NLR, and DOR were 27.0 (95% CI 6.3-116.4), 0.09 (95% CI 0.05-0.18), and 290 (95% CI 41-2046) for AI models, and 132.9 (95% CI 55.2-320.0), 0.04 (95% CI 0.01-0.09), and 3722 (95% CI 1634-8476) for human experts. The Deeks funnel plot demonstrated no significant publication bias for humans (P=.14; Figure S48 in Multimedia Appendix 1) and for AI models (P=.17; Figure S49 in Multimedia Appendix 1). To enable a direct human–AI comparison, we restricted the analysis to 4 matched test sets from 3 studies (Table S4 in Multimedia Appendix 1).

The PPV and NPV for "human experts" and "AI models" are provided in Figures S50 and S51 in Multimedia Appendix 1.

Publication Bias and Heterogeneity Analysis

In the overall analysis, the Deeks funnel plot asymmetry test (P=.65; Figure S52 in Multimedia Appendix 1) and the Egger test based on log(DOR) (P=.1875) detected no statistically significant asymmetry. Trim-and-fill and contour-enhanced funnel plots were also generated to further assess potential publication bias (Figures S53 and S54 in Multimedia Appendix 1).

In the overall analysis, substantial between-study heterogeneity was observed for both sensitivity and specificity, with I^2 values of 97.70% and 99.19%, respectively (both P<.001; Figure S55 in Multimedia Appendix 1). To explore potential sources of this heterogeneity, we conducted subgroup analyses (Figures S56-S59 in Multimedia Appendix 1) and performed meta-regression analysis with these covariates (Table S5 in Multimedia Appendix 1).

RoB Assessment

All 62 studies were assessed using QUADAS-2 (Figure S60 in Multimedia Appendix 1). Patient selection had the most judgments of high or unclear RoB (7 high and 18 unclear). These judgments mainly reflected insufficient reporting of case sources, unclear consecutive or random enrollment, or selective inclusion of cases with high-quality images or particular disease stages. The index-test RoB was high in 3 studies and unclear in 14, mainly because threshold prespecification or the independence of training and test data was insufficiently reported. Reference-standard RoB was unclear in 10 studies, largely because blinding or diagnostic criteria were not clearly described. Flow and timing were rated high risk in 5 studies and unclear in 6, mainly because the use of a consistent reference standard or assessment intervals was insufficiently reported.

Applicability concerns were generally low. For the patient selection, 1 study was rated as having high concern and 6 as unclear; for the index test, 1 was rated as having great concern and 2 as unclear. All 62 studies had low applicability concerns for the reference standard.

GRADE Assessment

The certainty of evidence for the diagnostic accuracy estimates from 62 studies was rated as moderate (Tables 4 and 5). The evidence was downgraded for serious inconsistency across studies. RoB, indirectness, and imprecision were not judged to be serious, and publication bias was not detected.

Table 4. GRADE[a] assessment and summary of findings for diagnostic accuracy.

Study profile			Results per 100 images	
Outcome	Studies (n)	Study design	Sensitivity, n (95% CI)	Specificity, n (95% CI)
True positives	62	Diagnostic accuracy studies	96 (94-97)	96 (94-97)
False negatives	62	Diagnostic accuracy studies	4 (3-6)	4 (3-6)

[a]GRADE: Grading of Recommendations Assessment, Development, and Evaluation.

Table 5. GRADE[a] assessment and summary of findings for diagnostic accuracy.

Study profile			Number of results per 100 images (95% CI)		
Outcome	Studies (n)	Study design	Sensitivity, n (95% CI)		Specificity, n (95% CI)
Not serious	Not serious	Serious	Not serious	Undetected	Moderate

[a]GRADE: Grading of Recommendations Assessment, Development, and Evaluation.

XSL•FO
RenderX

Discussion

Summary of the Main Findings

This systematic review and meta-analysis evaluated the diagnostic performance of image-based AI for ROP detection and severity classification. In subgroup analyses, diagnostic performance was generally favorable across most specific tasks. For TR-ROP, sensitivity remained high (0.97), whereas specificity was lower (0.74). In external validation, sensitivity and specificity were 0.95 and 0.83, respectively. These findings indicate that, despite the high accuracy of image recognition–based AI in diagnosing ROP, performance varied across diagnostic tasks and validation methods. Improving specificity for clinically important tasks and achieving stable performance across external datasets remain key challenges for future research.

Comparison With Other Previous Reviews

A systematic search for pertinent papers identified 3 systematic reviews with meta-analysis of image-based AI for ROP published before our search cutoff. Zhang et al [11] searched through June 13, 2021, and included 9 diagnostic accuracy studies comprising 15 DL classifiers. They separately pooled sensitivity and specificity for ROP detection (0.956 and 0.979) and differentiation between ROP grades (0.931 and 0.856) [11]. Chu et al [12] searched through October 1, 2022, and included 22 studies of image-based ML algorithms. They separately pooled sensitivity and specificity for ROP detection (0.93 and 0.95) and clinical subtype classification (0.93 and 0.93) and compared algorithm performance with that of clinical experts. Liu et al [87] searched studies published from January 2018 to July 2024 and included 14 studies. They separately pooled sensitivity and specificity for ROP detection (0.95 and 0.97) and plus disease detection (0.92 and 0.95). These reviews support the potential diagnostic value of image-based AI for ROP, while our research has examined finer disease classifications and a broader range of potential clinical applications. Extending the search through October 2026, our review provides pooled estimates for specific tasks, including TR-ROP, preplus and plus disease, and stage comparisons. We also stratified accuracy by validation methods and by whether TL was applied. These analyses place diagnostic accuracy in the context of the clinical task and validation setting, providing a clearer basis for developing and evaluating AI tools for ROP.

Discussion of the Main Findings

One important finding is that image-based AI showed strong discrimination not only between normal eyes and those with ROP but also in several subtype and severity tasks, including plus disease, preplus disease, TR-ROP, and pairwise stage classification. For the normal-vs-ROP task, high pooled sensitivity and specificity suggest potential value as an adjunct to routine screening, consistent with previous evidence. High AUCs for normal-vs-plus disease and pairwise stage comparisons further suggest that these models can identify imaging patterns associated with ROP severity. However, potential overlap in source datasets across studies may make these AUC estimates unstable. Specificity was lower despite high sensitivity in the normal-vs-TR-ROP and mild-vs-severe ROP tasks. Determining treatment needs requires assessment of disease zone, stage, and plus disease, making it more complex than detecting ROP alone [4]. Thresholds chosen to preserve sensitivity may also contribute to FP classifications. Although prioritizing sensitivity may be appropriate for initial screening, FP could increase the specialist review burden. Positive AI classifications should therefore be assessed by an ophthalmologist before treatment decisions are made.

Subgroup analyses by validation methods further suggested that model performance remained stable in internal validation and cross-validation, whereas pooled specificity declined in external validation. The external-validation subgroup maintained high sensitivity and AUC, indicating that the models retained some stability in identifying positive cases on independent datasets. However, the lower specificity suggests that their ability to correctly classify negative cases may be affected by differences between datasets. Variation in imaging devices, image quality, disease spectrum, annotation procedures, and reference standards across centers may contribute to reduced model performance [88]. If used for screening, the lower specificity could also increase FP referrals and the burden of specialist review. Therefore, although current findings support the potential utility of AI models in ROP image analysis, more multicenter, prospective studies with independent external validation are required before broader clinical implementation [89].

The subgroup analyses by model types indicated that DL algorithms achieved slightly higher pooled sensitivity, specificity, and AUC than traditional ML algorithms, suggesting advantages of DL in extracting complex fundus image features. Traditional ML largely depends on manually predefined or shallow features, while end-to-end DL training enables automatic learning of vascular morphology, lesion extent, and global image patterns [90]. However, whether the choice between DL and ML affects ROP detection performance remains uncertain, as the joint meta-regression test did not establish a difference between them (P=.08). We found that both TL-applied and no TL–applied subgroups had high diagnostic accuracy, with the former exhibiting slightly higher specificity and AUC. This result indicates that given limited ROP data sizes and substantial annotation costs, TL may improve model training efficiency and negative case identification capacity [42]. Nonetheless, its advantages require further verification using unified datasets and external validation frameworks.

This study also examined studies that compared AI models with ophthalmologists using the same datasets. In matched comparisons, both AI and ophthalmologists showed high overall diagnostic performance for ROP, with AI approaching the performance of ophthalmologists. AI showed higher sensitivity but lower specificity for ROP detection, suggesting fewer missed cases at the cost of more FP results. However, its sensitivity and specificity were lower for mild ROP/severe ROP, indicating that its performance in these more challenging diagnostic tasks requires further improvement. Given the limited number of comparable studies, this difference should be interpreted cautiously. A previous systematic review concluded that current evidence supports AI as an adjunct to clinical assessment but is insufficient to support its use as a standalone diagnostic tool for ROP [13]. In settings with limited ROP screening resources

XSL•FO
RenderX

and uneven access to specialists, a study using smartphone-based retinal imaging suggests that AI may help expand access to screening [53]. AI models may therefore support initial screening, risk stratification, and referral decisions. Whether their use improves screening efficiency and reduces missed diagnoses in clinical practice remains to be established in prospective studies.

Limitations

This study has several limitations, many of which could be addressed through higher-quality, more rigorously designed research. First, 51 included studies explicitly used a retrospective design, whereas prospective evidence and real-world clinical validation remained limited. This may increase the risk of case selection bias and affect the clinical applicability of the findings. Second, this review included only peer-reviewed publications in English. Studies reporting negative findings or lower diagnostic performance may therefore have been missed, creating a potential risk of publication bias. Where fewer than 10 studies were included in an analysis, the Deeks funnel plot asymmetry test had limited power.

Differences in imaging protocols and devices, image preprocessing, and image-labeling procedures may have affected comparability across studies. Variation in reference standards, patient characteristics and disease severity, model architecture, training strategies, and validation methods may also have contributed to between-study heterogeneity. Together with unequal sample sizes, these factors may have reduced the robustness of the pooled estimates. Validation of ROP models across populations and camera systems has also shown that differences in test data can affect model performance [88]. To investigate heterogeneity, we performed subgroup analyses by diagnostic task, model type, validation method, and use of TL. We also included these factors as covariates in meta-regression (Table S5 in Multimedia Appendix 1). The results suggest that diagnostic task, validation method, and use of TL may be associated with differences in AI diagnostic performance. However, heterogeneity remained high within subgroups, so these exploratory analyses did not fully explain the between-study differences. Moreover, many studies relied primarily on internal validation, and independent external validation was insufficient, limiting the generalizability of our findings to real-world clinical settings. Results from subgroups containing few studies should also be interpreted with caution.

Only 3 studies directly compared the diagnostic performance of AI models and human experts using the same datasets. Differences in experts' clinical experience and health care settings may have contributed to heterogeneity in their assessments; comparisons between AI and experts should therefore be interpreted with caution [91]. Moreover, this review primarily synthesized diagnostic accuracy measures, such as sensitivity and specificity, but cannot determine whether integrating AI into routine screening would reduce missed diagnoses, ease the burden of referrals and specialist review, or improve the timeliness of treatment.

Conclusions

Image-based AI models achieve high overall diagnostic accuracy in ROP detection and severity classification, although their specificity for identifying TR-ROP and in external validation remains to be improved. Current evidence supports their use as adjuncts to screening and risk stratification, with clinical experts involved in key diagnostic and treatment decisions. Future studies should standardize image-quality criteria and reference standards, conduct multicenter prospective external validation across countries, imaging devices, and health care settings with different levels of resources, and more fully report model development procedures and failure cases.

Acknowledgments

We acknowledge the contributions of the researchers whose work provided the foundation for this systematic review and meta-analysis.

Data Availability

The original contributions presented in the study are included in the papers further inquiries can be directed to the corresponding author.

Funding

This work was supported by the Application Basic Research Project of Sichuan Provincial Department of Science and Technology (2020YJ0448). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of this paper.

Generative AI Use

The generative AI tool ChatGPT (GPT-5.6; OpenAI) was used solely for English language editing, grammatical correction, and refinement of academic expression to improve the clarity and readability of this paper. All AI-assisted language revisions were reviewed and verified by the authors, who take full responsibility for the accuracy, integrity, and scientific content of the final paper.

Authors' Contributions

Conceptualization: HB, YZ, YB, YW, YY

XSL•FO
RenderX

Formal analysis: HB, YZ
Funding acquisition: YY
Investigation: HB, YZ
Methodology: HB, YZ, YB, YW, YY
Resources: YY
Supervision: YY
Writing–original draft preparation: HB, YZ
Writing–review and editing: HB, YZ, YY
All authors commented on previous versions of the paper. All authors read and approved this final paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Extensive materials on the study.
[PDF File (Adobe PDF File), 33923 KB-Multimedia Appendix 1]

Multimedia Appendix 2

PRISMA-DTA checklist.
[PDF File (Adobe PDF File), 97 KB-Multimedia Appendix 2]

References

1. Sabri K, Ells AL, Lee EY, Dutta S, Vinekar A. Retinopathy of prematurity: a global perspective and recent developments. Pediatrics. 2022;150(3):e2021053924. [doi: 10.1542/peds.2021-053924] [Medline: 35948728]
2. Early Treatment for Retinopathy of Prematurity Cooperative Group. Revised indications for the treatment of retinopathy of prematurity: results of the early treatment for retinopathy of prematurity randomized trial. Arch Ophthalmol. 2003;121(12):1684-1694. [doi: 10.1001/archopht.121.12.1684] [Medline: 14662586]
3. Biten H, Redd TK, Moleta C, Campbell JP, Ostmo S, Jonas K, et al. Diagnostic accuracy of ophthalmoscopy vs telemedicine in examinations for retinopathy of prematurity. JAMA Ophthalmol. 2018;136(5):498-504. [FREE Full text] [doi: 10.1001/jamaophthalmol.2018.0649] [Medline: 29621387]
4. Chiang MF, Quinn GE, Fielder AR, Ostmo SR, Paul CRV, Berrocal A. International classification of retinopathy of prematurity, third edition. Ophthalmology. 2021;128(10):e51-e68. [doi: 10.1016/j.ophtha.2021.05.031] [Medline: 34247850]
5. Fierson WM, American Academy of Pediatrics Section on Ophthalmology, American Academy of Ophthalmology, American Association for Pediatric Ophthalmology and Strabismus, American Association of Certified Orthoptists. Screening examination of premature infants for retinopathy of prematurity. Pediatrics. Dec 2018;142(6):e20183810. [doi: 10.1542/peds.2018-3061] [Medline: 30478242]
6. Nakayama LF, Mitchell WG, Ribeiro LZ, Dychiao RG, Phanphruk W, Celi LA, et al. Fairness and generalisability in deep learning of retinopathy of prematurity screening algorithms: a literature review. BMJ Open Ophthalmol. 2023;8(1):e001216. [FREE Full text] [doi: 10.1136/bmjophth-2022-001216] [Medline: 37558406]
7. Chiang MF, Jiang L, Gelman R, Du YE, Flynn JT. Interexpert agreement of plus disease diagnosis in retinopathy of prematurity. Arch Ophthalmol. 2007;125(7):875-880. [doi: 10.1001/archopht.125.7.875] [Medline: 17620564]
8. Jafarizadeh A, Maleki SF, Pouya P, Sobhi N, Abdollahi M, Pedrammehr S, et al. Current and future roles of artificial intelligence in retinopathy of prematurity. Artif Intell Rev. 2025;58(6):188. [doi: 10.1007/s10462-025-11153-6]
9. Brown JM, Campbell JP, Beers A, Chang K, Ostmo S, Chan RVP, et al. Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. JAMA Ophthalmol. 2018;136(7):803-810. [FREE Full text] [doi: 10.1001/jamaophthalmol.2018.1934] [Medline: 29801159]
10. Wagner SK, Liefers B, Radia M, Zhang G, Struyven R, Faes L, et al. Development and international validation of custom-engineered and code-free deep-learning models for detection of plus disease in retinopathy of prematurity: a retrospective study. Lancet Digit Health. 2023;5(6):e340-e349. [FREE Full text] [doi: 10.1016/S2589-7500(23)00050-X] [Medline: 37088692]
11. Zhang J, Liu Y, Mitsuhashi T, Matsuo T. Accuracy of deep learning algorithms for the diagnosis of retinopathy of prematurity by fundus images: a systematic review and meta-analysis. J Ophthalmol. 2021;2021:8883946. [FREE Full text] [doi: 10.1155/2021/8883946] [Medline: 34394982]
12. Chu Y, Hu S, Li Z, Yang X, Liu H, Yi X, et al. Image analysis-based machine learning for the diagnosis of retinopathy of prematurity: a meta-analysis and systematic review. Ophthalmol Retina. 2024;8(7):678-687. [doi: 10.1016/j.oret.2024.01.013] [Medline: 38237772]

XSL•FO
RenderX

13. Ramanathan A, Athikarisamy SE, Lam GC. Artificial intelligence for the diagnosis of retinopathy of prematurity: a systematic review of current algorithms. Eye (Lond). 2023;37(12):2518-2526. [doi: 10.1038/s41433-022-02366-y] [Medline: 36577806]
14. Mutua EN, Kasamani BS, Reich C. Deep learning applications for diabetic retinopathy and retinopathy of prematurity diseases diagnosis: a systematic review. Int J Ophthalmol. 2025;18(8):1594-1602. [doi: 10.18240/ijo.2025.08.23] [Medline: 40827296]
15. Zhao X, Wu Z, Wu S, Cui K, Lam W, Wu W, et al. Multimodal artificial intelligence in retinopathy of prematurity: a comprehensive narrative review. Surv Ophthalmol. 2026;71(6):1539-1557. [doi: 10.1016/j.survophthal.2026.05.006] [Medline: 42134610]
16. Bai A, Carty C, Dai S. Performance of deep-learning artificial intelligence algorithms in detecting retinopathy of prematurity: a systematic review. Saudi J Ophthalmol. 2022;36(3):296-307. [FREE Full text] [doi: 10.4103/sjopt.sjopt_219_21] [Medline: 36276252]
17. Bai A, Dai S, Hung J, Kirpalani A, Russell H, Elder J, et al. Multicenter validation of deep learning algorithm ROP.AI for the automated diagnosis of plus disease in ROP. Transl Vis Sci Technol. 2023;12(8):13. [FREE Full text] [doi: 10.1167/tvst.12.8.13] [Medline: 37578427]
18. Salameh JP, Bossuyt PM, McGrath TA, Thombs BD, Hyde CJ, Macaskill P, et al. Preferred Reporting Items for Systematic Review and Meta-Analysis of Diagnostic Test Accuracy Studies (PRISMA-DTA): explanation, elaboration, and checklist. BMJ. 2020;370:m2632. [FREE Full text] [doi: 10.1136/bmj.m2632] [Medline: 32816740]
19. Hu J, Chen Y, Zhong J, Ju R, Yi Z. Automated analysis for retinopathy of prematurity by deep neural networks. IEEE Trans Med Imaging. 2019;38(1):269-279. [doi: 10.1109/TMI.2018.2863562] [Medline: 30080144]
20. Tan Z, Simkin S, Lai C, Dai S. Deep learning algorithm for automated diagnosis of retinopathy of prematurity plus disease. Transl Vis Sci Technol. 2019;8(6):23. [FREE Full text] [doi: 10.1167/tvst.8.6.23] [Medline: 31819832]
21. Wang Y, Chen Y. Automated recognition of retinopathy of prematurity with deep neural networks. J Phys: Conf Ser. 2019;1187:042057. [FREE Full text] [doi: 10.1088/1742-6596/1187/4/042057]
22. Huang YP, Vadloori S, Chu HC, Kang EYC, Wu WC, Kusaka S, et al. Deep learning models for automated diagnosis of retinopathy of prematurity in preterm infants. Electronics. 2020;9(9):1444. [doi: 10.3390/electronics9091444]
23. Zhang Y, Wang L, Wu Z, Zeng J, Chen Y, Tian R, et al. Development of an automated screening system for retinopathy of prematurity using a deep neural network for wide-angle retinal images. IEEE Access. 2019;7:10232-10241. [doi: 10.1109/access.2018.2881042]
24. Wang J, Ji J, Zhang MZ, Lin JW, Zhang GH, Gong WF, et al. Automated explainable multidimensional deep learning platform of retinal images for retinopathy of prematurity screening. JAMA Netw Open. May 05, 2021;4(5):e218758. [doi: 10.1001/jamanetworkopen.2021.8758] [Medline: 33950206]
25. Cole E, Valikodath NG, Al-Khaled T, Bajimaya S, Kc S, Chuluunbat T, et al. Evaluation of an artificial intelligence system for retinopathy of prematurity screening in Nepal and Mongolia. Ophthalmol Sci. 2022;2(4):100165. [FREE Full text] [doi: 10.1016/j.xops.2022.100165] [Medline: 36531583]
26. Coyner AS, Oh MA, Shah PK, Singh P, Ostmo S, Valikodath NG, et al. External validation of a retinopathy of prematurity screening model using artificial intelligence in 3 low- and middle-income populations. JAMA Ophthalmol. 2022;140(8):791-798. [FREE Full text] [doi: 10.1001/jamaophthalmol.2022.2135] [Medline: 35797036]
27. Li P, Liu J. Early diagnosis and quantitative analysis of stages in retinopathy of prematurity based on deep convolutional neural networks. Transl Vis Sci Technol. 2022;11(5):17. [FREE Full text] [doi: 10.1167/tvst.11.5.17] [Medline: 35579887]
28. Luo Z, Ding X, Hou N, Wan J. A deep-learning-based collaborative edge-cloud telemedicine system for retinopathy of prematurity. Sensors (Basel). 2022;23(1):276. [FREE Full text] [doi: 10.3390/s23010276] [Medline: 36616874]
29. Wu Q, Hu Y, Mo Z, Wu R, Zhang X, Yang Y, et al. Development and validation of a deep learning model to predict the occurrence and severity of retinopathy of prematurity. JAMA Netw Open. 2022;5(6):e2217447. [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.17447] [Medline: 35708686]
30. Attallah O. GabROP: gabor wavelets-based CAD for retinopathy of prematurity diagnosis via convolutional neural networks. Diagnostics (Basel). 2023;13(2):171. [FREE Full text] [doi: 10.3390/diagnostics13020171] [Medline: 36672981]
31. Brindha R, Lakkshmanan A, Renukadevi P, Jeyakumar D. Detection of retinopathy of prematurity using ResNet based deep features and support vector machine classifier. 2023. Presented at: 2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS); August 23-25, 2023:1020-1027; Trichy, India. [doi: 10.1109/icaiss58487.2023.10250647]
32. Krishna MV, Rao BS. An approach to classify and segment diabetic retinopathy and retinopathy of prematurity. Trait Signal. 2023;40(3):939-948. [doi: 10.18280/ts.400310]
33. Mutua EN, Kasamani BS, Reich C. Smartphone fundus photography enhancement for retinopathy of prematurity disease diagnosis using deep learning. 2023. Presented at: 2023 6th International Conference of Computer and Informatics Engineering (IC2IE); September 14-15, 2023; Lombok, Indonesia. [doi: 10.1109/ic2ie60547.2023.10331154]
34. Ramachandran S. A novel deep learning framework for the identification of tortuous vessels in plus diseased infant retinal images. Intell Data Anal. 2023;27(5):1523-1537. [doi: 10.3233/ida-220451]

XSL•FO
RenderX

35. Rao DP, Savoy FM, Tan JZE, Fung BP, Bopitiya CM, Sivaraman A, et al. Development and validation of an artificial intelligence based screening tool for detection of retinopathy of prematurity in a South Indian population. Front Pediatr. 2023;11:1197237. [FREE Full text] [doi: 10.3389/fped.2023.1197237] [Medline: 37794964]
36. Sankari VMR, Umapathy U, Alasmari S, Aslam SM. Automated detection of retinopathy of prematurity using quantum machine learning and deep learning techniques. IEEE Access. 2023;11:94306-94321. [doi: 10.1109/access.2023.3311346]
37. Shen Y, Luo Z, Xu M, Liang Z, Fan X, Lu X. Automated detection for retinopathy of prematurity with knowledge distilling from multi-stream fusion network. Knowledge-Based Syst. 2023;269:110461. [doi: 10.1016/j.knosys.2023.110461]
38. Subramaniam A, Orge F, Douglass M, Can B, Monteoliva G, Fried E, et al. Image harmonization and deep learning automated classification of plus disease in retinopathy of prematurity. J Med Imaging (Bellingham). 2023;10(6):061107. [doi: 10.1117/1.JMI.10.6.061107] [Medline: 37794884]
39. Young BK, Cole ED, Shah PK, Ostmo S, Subramaniam P, Venkatapathy N, et al. Efficacy of smartphone-based telescreening for retinopathy of prematurity with and without artificial intelligence in India. JAMA Ophthalmol. 2023;141(6):582-588. [FREE Full text] [doi: 10.1001/jamaophthalmol.2023.1466] [Medline: 37166816]
40. Almeida J, Kubicek J, Penhaker M, Cerny M, Augustynek M, Varysova A, et al. Enhancing ROP plus form diagnosis: an automatic blood vessel segmentation approach for newborn fundus images. Results Eng. 2024;24:103054. [doi: 10.1016/j.rineng.2024.103054]
41. Liu Y, Xie H, Zhao X, Tang J, Yu Z, Wu Z, et al. Automated detection of nine infantile fundus diseases and conditions in retinal images using a deep learning system. EPMA J. 2024;15(1):39-51. [doi: 10.1007/s13167-024-00350-y] [Medline: 38463622]
42. Feng W, Huang Q, Ma T, Ju L, Ge Z, Chen Y, et al. Development and validation of a semi-supervised deep learning model for automatic retinopathy of prematurity staging. iScience. 2024;27(1):108516. [FREE Full text] [doi: 10.1016/j.isci.2023.108516] [Medline: 38269093]
43. Selvamarilakshmi D, Vani R, Aravinth S, Thenappan S, Begam R, Gangarani N. Retracted: integrating machine learning for early diagnosis and prognostic assessment of retinopathy of premature babies. 2024. Presented at: 2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI); May 09-10, 2024:1-7; Chennai, India. [doi: 10.1109/accai61061.2024.10602334]
44. Cai R, Li J, Wang F, Li X, Guo J, Dai J, et al. A hybrid swin transformer and EfficientNetB2-based framework for automated ROP detection in premature infants. Int J Imaging Syst Techol. 2025;35(5):e70191. [doi: 10.1002/ima.70191]
45. Chen S, Chen F, Huang Z, Gu Y, Zhang G. Diagnosis system for retinopathy of prematurity with Fourier parameterized rotation equivariant convolutions network and prompt mechanism. Expert Syst Appl. 2025;286:128069. [doi: 10.1016/j.eswa.2025.128069]
46. Deepthi K, Josephine MS, Jayabala Raja V. Automated diagnosis of plus disease in retinopathy of prematurity based on transformer-based unsupervised curriculum learning. Biomed Signal Process Control. 2025;104:107521. [doi: 10.1016/j.bspc.2025.107521]
47. Engin CD, Ozturk T, Ozkan O, Oztas A, Selver MA, Tuzun F. Prediction of retinopathy of prematurity development and treatment need with machine learning models. BMC Ophthalmol. 2025;25(1):194. [FREE Full text] [doi: 10.1186/s12886-025-04025-8] [Medline: 40211206]
48. Kaushik J, Khalid SB, Vittal S, Muralidhar S, Vinekar A, Srinivasa G. Automated detection of plus disease from infant retinal images. 2025. Presented at: 2025 15th International Conference on Electrical Engineering (ICEENG); May 12-15, 2025:1-6; Cairo, Egypt. [doi: 10.1109/iceeng64546.2025.11031393]
49. Kariakin M, Stepanova E, Kashkin A, Yakutin S, Malakhov E, Reshetnikov K. Automated classification of prematurity retinopathy using ViT, resNet, and efficientNet. 2025. Presented at: 2025 IEEE Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT); May 12-13, 2025:65-68; Yekaterinburg, Russian Federation. [doi: 10.1109/usbereit65494.2025.11054213]
50. Krishna SS, Reka SS. Early detection of retinopathy of prematurity using voting classifier-based ensemble deep learning models. Int J Comput Intell Syst. 2025;18(1):238. [doi: 10.1007/s44196-025-00847-y]
51. Mahrous F, Abu-Leif H, Modyiq T, AlGhamdi H, AlSaedi K, Algamdi A. Towards privacy-preserving big data analysis with decentralized learning for ROP imaging. 2025. Presented at: 2025 International Conference on Innovation in Artificial Intelligence and Internet of Things (AIIT); May 07-08, 2025:1-10; Jeddah, Saudi Arabia. [doi: 10.1109/aiit63112.2025.11082911]
52. Mowla N, Mowla M, Rabie K, Alsinglawi B. A lightweight deep learning model for retinopathy of prematurity classification in eHealth applications. 2025. Presented at: 2025 International Wireless Communications and Mobile Computing (IWCMC); May 12-16, 2025:227-232; Abu Dhabi, United Arab Emirates. [doi: 10.1109/iwcmc65282.2025.11059591]
53. Ortiz A, Patiño S, Torres J, Mármol J, Serafin C, Dodhia R, et al. AI-enabled screening for retinopathy of prematurity in low-resource settings. JAMA Netw Open. 2025;8(4):e257831. [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.7831] [Medline: 40299381]
54. Rahim S, Sabri K, Ells A, Wayssyng A, Lawford M, Shu L. Using novel fundus image preprocessing to improve the classification of retinopathy of prematurity (ROP) using deep learning. 2025 International Conference on Intelligent Systems.

XSL•FO
RenderX

Theories and Applications (SITA); 2025. Presented at: 2025 International Conference on Intelligent Systems: Theories and Applications (SITA); October 20-21, 2025:1-8; Rabat, Morocco. [doi: 10.1109/sita67914.2025.11273702]
55. Varysova A, Kubicek J, Penhaker M, Augustynek M, Oczka D, Marsolkova K, et al. Modeling and recognition of retinal blood vessels tortuosity in ROP plus disease: a hybrid segmentation-classification scheme. Int J Intell Syst. 2025;2025:6688133. [doi: 10.1155/int/6688133]
56. Wang D, Qiao W, Guo W, Cai Y, Liu M. Diagnosis of retinopathy of prematurity based on deep learning framework with SK attention mechanism. 2025. Presented at: 2025 IEEE International Conference on Real-time Computing and Robotics (RCAR); June 01-06, 2025:1247-1252; Toyama, Japan. [doi: 10.1109/RCAR65431.2025.11139554]
57. J K, Bin Khalid S, Vittal S, Muralidhar S, Vinekar A, Srinivasa G. Automated diagnosis of early-stage retinopathy of prematurity. Eye (Lond). Jan 2026;40(1):134-138. [doi: 10.1038/s41433-025-04122-4] [Medline: 41286176]
58. Alam MM, Siddiqui A, Aslam Z, Kumar S, Samad A. Automating ROP diagnosis and severity with deep learning. 2025. Presented at: 2025 IEEE 22nd International Conference on Smart Communities: Improving Quality of Life using AI, Robotics and IoT (HONET); December 02-04, 2025; Topi, District Swabi, KP, Pakistan. [doi: 10.1109/honet67928.2025.11318433]
59. De D, Pal MN, Hazra D. QROP: Quantum learning ‐ based identification of retinopathy of prematurity. IET Quantum Commun. 2025;6(1):e70008. [doi: 10.1049/qtc2.70008]
60. He D, Luo X, Ying B, Quinn GE, Baumritter A, Chen Y, et al. Machine learning models for predicting treatment-requiring retinopathy of prematurity in the e-ROP study. Transl Vis Sci Technol. 2025;14(8):14. [FREE Full text] [doi: 10.1167/tvst.14.8.14] [Medline: 40787926]
61. Yurdakul M, Uyar K, Taşdemír Ş, Atabaş Í. ROPGCViT: a novel explainable vision transformer for retinopathy of prematurity diagnosis. IEEE Access. 2025;13:77064-77079. [doi: 10.1109/access.2025.3564213]
62. Shoeibi N, Ameri N, Hoseinkhani MR, Astaneh MRA, Shariati MM, Hosseini SM, et al. Deep learning algorithms for timely diagnosis of retinopathy of prematurity requiring treatment. Eye (Lond). 2026;40(1):63-70. [doi: 10.1038/s41433-025-04096-3] [Medline: 41184508]
63. Wang J, Ju R, Chen Y, Zhang L, Hu J, Wu Y, et al. Automated retinopathy of prematurity screening using deep neural networks. eBioMedicine. 2018;35:361-368. [FREE Full text] [doi: 10.1016/j.ebiom.2018.08.033] [Medline: 30166272]
64. Ding A, Chen Q, Cao Y, Liu B. Retinopathy of prematurity stage diagnosis using object segmentation and convolutional neural networks. 2020. Presented at: 2020 International Joint Conference on Neural Networks (IJCNN); July 19-24, 2020:1-6; Glasgow, United Kingdom. [doi: 10.1109/ijcnn48605.2020.9207288]
65. Coyner AS, Chen JS, Singh P, Schelonka RL, Jordan BK, McEvoy CT, et al. Single-examination risk prediction of severe retinopathy of prematurity. Pediatrics. 2021;148(6):e2021051772. [FREE Full text] [doi: 10.1542/peds.2021-051772] [Medline: 34814160]
66. Ramachandran S, Niyas P, Vinekar A, John R. A deep learning framework for the detection of Plus disease in retinal fundus images of preterm infants. Biocybern Biomed Eng. 2021;41(2):362-375. [doi: 10.1016/j.bbe.2021.02.005]
67. Agrawal R, Kulkarni S, Walambe R, Deshpande M, Kotecha K. Deep dive in retinal fundus image segmentation using deep learning for retinopathy of prematurity. Multimed Tools Appl. 2022;81(8):11441-11460. [doi: 10.1007/s11042-022-12396-z]
68. Zhang R, Zhao J, Xie H, Wang T, Chen G, Zhang G, et al. Automatic diagnosis for aggressive posterior retinopathy of prematurity via deep attentive convolutional neural network. Expert Syst Appl. 2022;187:115843. [doi: 10.1016/j.eswa.2021.115843]
69. Redd TK, Campbell JP, Brown JM, Kim SJ, Ostmo S, Chan RVP, et al. Evaluation of a deep learning image assessment system for detecting severe retinopathy of prematurity. Br J Ophthalmol. 2019;103:580-584. [FREE Full text] [doi: 10.1136/bjophthalmol-2018-313156] [Medline: 30470715]
70. Legocki AT, Lee AY, Ding L, Moshiri Y, Zepeda EM, Gillette TB, et al. Multivariate models to diagnose early referral-warranted retinopathy of prematurity with handheld optical coherence tomography. Transl Vis Sci Technol. 2023;12(5):26. [FREE Full text] [doi: 10.1167/tvst.12.5.26] [Medline: 37223917]
71. Agrawal R, Kulkarni S, Deshpande M, Gaikwad A, Walambe R, Kotecha KV. Classification of stages 1,2,3 and preplus, plus disease of ROP using MultiCNN_LSTM classifier. MethodsX. 2025;14:103182. [FREE Full text] [doi: 10.1016/j.mex.2025.103182] [Medline: 39944106]
72. Salih N, Ksantini M, Hussein N, Halima DB, Razzaq AA, Ahmed S. Deep learning models and fusion classification technique for accurate diagnosis of retinopathy of prematurity in preterm newborn. Baghdad Sci J. 2024;21(5):21. [doi: 10.21123/bsj.2023.8747]
73. Yang Y, Li L, Li Y, Song L, Zuo H, Shi R, et al. Construction of a predictive model for retinopathy of prematurity using machine learning algorithms. BMC Ophthalmol. 2025;25(1):531. [FREE Full text] [doi: 10.1186/s12886-025-04382-4] [Medline: 41034754]
74. Vahidmoghadam M, Ghorbani P, Ahmadi MJ, Khameneh EA, Tavassoli B, Taghirad HD. Automated diagnosis of plus form and early stages of ROP using deep learning models. Sci Rep. 2026;16(1):7234. [FREE Full text] [doi: 10.1038/s41598-026-37064-2] [Medline: 41639314]

XSL•FO
RenderX

75. Zhang X, Liang H, Wang R, Wu R, Shen X, Zhang Y. Research on the construction of an AI diagnostic model for plus disease of retinopathy of prematurity based on cross-center fusion datasets. Front Pediatr. 2026;14:1765353. [FREE Full text] [doi: 10.3389/fped.2026.1765353] [Medline: 41953179]
76. Zhao J, Li L, Lin Z, Luo M, Yun D, Wen J, et al. DeepAdapter: a generalisable algorithm integrating self-supervised learning and unsupervised domain adaptation for robust retinopathy of prematurity screening. Br J Ophthalmol. 2026. [FREE Full text] [doi: 10.1136/bjo-2025-328559] [Medline: 42285743]
77. Huang Y, Basanta H, Kang EY, Chen K, Hwang Y, Lai C, et al. Automated detection of early-stage ROP using a deep convolutional neural network. Br J Ophthalmol. 2021;105(8):1099-1103. [FREE Full text] [doi: 10.1136/bjophthalmol-2020-316526] [Medline: 32830123]
78. Derks LA, Tekin YS, Loudon SE, Vingerling JR, Coyner AS, Campbell JP, et al. External validation of an artificial intelligence–based model for retinopathy of prematurity screening using Phoenix ICON retinal images. JAAPOS. 2025;29(6):104696. [FREE Full text] [doi: 10.1016/j.jaapos.2025.104696] [Medline: 41276153]
79. Murugan R, Sekar K, Akshaya V, Deepthi RS, Harshini JP, Mahi AR. Deep learning-based detection of retinopathy of prematurity using vision transformers. 2025. Presented at: 2025 Second International Conference on Networks and Soft Computing (ICNSoC); June 12-14, 2025:1-6; Vadlamudi, Guntur, Andhra Pradesh, India. [doi: 10.1109/icnsoc66817.2025.00080]
80. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. Oct 18, 2011;155(8):529-536. [FREE Full text] [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
81. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37-46. [doi: 10.1177/001316446002000104]
82. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336(7650):924-926. [FREE Full text] [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]
83. Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol. 2005;58(10):982-990. [doi: 10.1016/j.jclinepi.2005.02.022] [Medline: 16168343]
84. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2024. URL: https://www.R-project.org/ [accessed 2026-09-09]
85. Campbell JP, Ataer-Cansizoglu E, Bolon-Canedo V, Bozkurt A, Erdogmus D, Kalpathy-Cramer J, et al. ImagingInformatics in ROP (i-ROP) Research Consortium. Expert diagnosis of plus disease in retinopathy of prematurity from computer-based image analysis. JAMA Ophthalmol. 2016;134(6):651-657. [FREE Full text] [doi: 10.1001/jamaophthalmol.2016.0611] [Medline: 27077667]
86. Bradshaw TJ, Huemann Z, Hu J, Rahmim A. A guide to cross-validation for artificial intelligence in medical imaging. Radiol Artif Intell. 2023;5(4):e220232. [FREE Full text] [doi: 10.1148/ryai.220232] [Medline: 37529208]
87. Liu R, Liu G, Jiang X, Lu F. Application and performance of artificial intelligence in screening retinopathy of prematurity from 2018 to 2024: a meta-analysis and systematic review. Eye Sci. 2014;1(3):206-223. [doi: 10.12419/es24082501]
88. Chen J, Coyner A, Ostmo S, Sonmez K, Bajimaya S, Pradhan E. Deep learning for the diagnosis of stage inRetinopathy of prematurity: accuracy and generalizability across populations and cameras. Ophthalmol Retina. 2021;5(10):1027-1035. [FREE Full text] [doi: 10.1016/j.oret.2020.12.013]
89. Sounderajah V, Guni A, Liu X, Collins GS, Karthikesalingam A, Markar SR, Ben Glocker, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. Nat Med. 2025;31(10):3283-3289. [doi: 10.1038/s41591-025-03953-8]
90. Ataer-Cansizoglu E, Bolon-Canedo V, Campbell JP, Bozkurt A, Erdogmus D, Kalpathy-Cramer J, et al. i-ROP Research Consortium. Computer-based image analysis for plus disease diagnosis in retinopathy of prematurity: performance of the "i-ROP" system and image features associated with expert diagnosis. Transl Vis Sci Technol. 2015;4(6):5. [FREE Full text] [doi: 10.1167/tvst.4.6.5] [Medline: 26644965]
91. Myung JS, Chan RVP, Espiritu MJ, Williams SL, Granet DB, Lee TC, et al. Accuracy of retinopathy of prematurity image-based diagnosis by pediatric ophthalmology fellows: implications for training. JAAPOS. 2011;15(6):573-578. [FREE Full text] [doi: 10.1016/j.jaapos.2011.06.011] [Medline: 22153403]

Abbreviations

AUC: area under the curve
DL: deep learning
DOR: diagnostic odds ratio
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
ICROP3: International Classification of Retinopathy of Prematurity, Third Edition
ML: machine learning

XSL•FO
RenderX

NLR: negative likelihood ratio
NPV: negative predictive value
PLR: positive likelihood ratio
PPV: positive predictive value
PRISMA-DTA: Preferred Reporting Items for a Systematic Review and Meta-Analysis of Diagnostic Test Accuracy Studies
PROBAST+AI: Prediction Model Risk of Bias Assessment Tool + Artificial Intelligence
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies-2
QUADAS-AI: Quality Assessment of Diagnostic Accuracy Studies–Artificial Intelligence
RoB: risk of bias
ROP: retinopathy of prematurity
SROC: summary receiver operating characteristic
TL: transfer learning
TR-ROP: treatment-requiring retinopathy of prematurity

Edited by I Steenstra; submitted 09.Aug.2026; peer-reviewed by SH Teshnizi, A Jafarizadeh; comments to author 01.Oct.2026; revised version received 29.Oct.2026; accepted 30.Oct.2026; published 19.Nov.2026

Please cite as:
Bai H, Zhou Y, Bai Y, Wang Y, Yang Y
Image-Based AI for the Diagnosis of Retinopathy of Prematurity: Systematic Review and Meta-Analysis
JMIR Med Inform 2026;14:e109141
URL: https://medinform.jmir.org/2026/1/e109141
doi: 10.2196/109141
PMID:

©Haiyao Bai, Yuran Zhou, Yang Bai, Yaolu Wang, Ying Yang. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 19.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX