International Journal of Retina (*IJRETINA*) 2025, Volume 8, Number 2.
P-ISSN. 2614-8684, E-ISSN.2614-8536

THE USE OF ARTIFICIAL INTELLIGENCE FOR DIAGNOSING RETINOPATHY OF PREMATURITY A SYSTEMATIC REVIEW

Adinda Mulya Pertiwi[1], Yeni Dwi Lestari[2], Dian Estu Yulia[3]*

[1]*Residency Program in Ophthalmology, Department of Ophthalmology, Faculty of Medicine Universitas Indonesia – Cipto Mangunkusumo General Hospital, Jakarta, Indonesia*

[2] *Community Ophthalmology Division, Department of Ophthalmology, Faculty of Medicine Universitas Indonesia – Cipto Mangunkusumo General Hospital, Jakarta, Indonesia*

[3] *Pediatric Ophthalmology Division, Department of Ophthalmology, Faculty of Medicine Universitas Indonesia – Cipto Mangunkusumo General Hospital, Jakarta, Indonesia*

Abstract

***Introduction*:** Retinopathy of prematurity (ROP) is a major but preventable cause of childhood blindness. Screening in developing countries is challenging due to skilled staff shortages. Recent advances in artificial intelligence (AI) offer promising result. This study evaluates the diagnostic performance of AI models for ROP screening.

Methods: This systematic review followed PRISMA guidelines and included studies from Cochrane, MEDLINE, and ScienceDirect. Eligible studies were cross-sectional or cohort designs that compared AI diagnostic accuracy for ROP against a gold standard and reported relevant metrics. Studies were graded using the Oxford CEBM levels of evidence.

***Results*:** Of 608 studies, 12 were included. i-ROP DL showed high sensitivity and specificity (AUC ~0.99), with ResNet-152 and EfficientNet-B0 also performing well. Despite variations in specificity and PPV, AI shows promise for ROP screening. i-ROP DL and ResNet-152 may need demographic adaptation. Though cost-effectiveness data are lacking, AI could reduce workload and improve diagnostic consistency.

Conclusion: AI shows high sensitivity, but variable specificity highlights the need for refinement. The review also underscores the importance of validation across diverse populations to ensure generalizability. AI integration in clinical practice can enhance early detection, standardize diagnoses, and alleviate the burden on healthcare professionals, particularly in low-resource settings.

Keywords: Artificial Intelligence, Retinopathy of Prematurity. **Cite This Article:** PERTIWI, Adinda Mulya; YULIA, Dian Estu; LESTARI, Yeni Dwi. The use of Artificial Intelligence for Diagnosing Retinopathy of Prematurity – A Systematic Review. **International Journal of Retina**, [S.l.], v. 8, n. 2, p. 123, oct. 2025. ISSN 2614-8536. Available at: <https://www.ijretina.com/index.php/ijretina/article/view/316>. Date accessed: 01 oct. 2025. doi: https://doi.org/10.35479/ijretina.2025.vol008.iss002.316....

Correspondence to:
Adinda Mulya Pertiwi,
Universitas Indonesia – Cipto Mangunkusumo General Hospital, Jakarta, Indonesia,
adinda.mulya@yahoo.com

INTRODUCTION

Retinopathy of prematurity (ROP) is a vaso-proliferative disease of the retina associated with prematurity and the leading cause of childhood blindness worldwide.[1] A multicenter analysis of Early Treatment for Retinopathy of Prematurity (ET-ROP) showed that 68% of premature infants with less than 1250 gram of bodyweight will develop at least mild ROP. A multicenter study in Indonesia, the incidence of all-stage ROP was 18% and in Cipto Mangunkusumo Hospital was 4.8% in 2014.[3] It is estimated that more than 10% of premature infants with ROP will develop severe visual impairment and blindness.[4] Global burden of disease analysis showed that in 2010, there were estimated around 257,000 years lived with disability due to visual impairment associated with ROP.[5] The underlying link between prematurity and development of this disease is because the nasal and temporal portions of the retina form late in pregnancy, 32 and 40 weeks respectively causing preterm birth infants had less developed retina.[1] Birth body weight is also known to strongly associated with ROP.[1]

Guidelines from the American Association for Pediatric Ophthalmology and Strabismus, American Academy of Pediatrics, and American Academy of Ophthalmology state that infants born ≤30 weeks gestational age or ≤1500 gram of body weight is a candidate for screening.[6] Screening for ROP requires bedside or telemedicine examination of fundus image. Screening for ROP in Indonesia is also done in some hospitals, especially those in big cities. The screening criteria in Indonesia refer to the recommendations from the 2014 RoP national *Pokja* and Premature Infant Working Group workshop. These criteria also use references from the United States. Screening is carried out on babies with a birth weight of <1500 grams or a gestational age of <34 weeks, or babies with risk factors. In India, the screening criteria for Retinopathy of Prematurity also refer to the same criteria as in Indonesia. Several ROP screening programs are conducted like a multicenter study conducted at Harapan Kita Women and Children's Health Centre and Cipto Mangunkusumo Hospital.[3] Jakarta-ROP (JakROP) is one of Cipto Mangunkusumo Hospital's flagship mobile ROP screening program in several selected vertical hospital in Jakara. In a general population, only 5-10% of babies screened will develop visual impairment secondary to ROP. However, there are a number of challenges for this screening. Regular and wide population screening is difficult especially in low- and middle-income countries usually associated with inadequate training, remote area, and skilled staff shortages.[8] This lacked of skilled staff, especially physician that able to recognized and diagnosed ROP from fundus image is the core problem tackled by many in rural developing regions/countries. Another challenge to ROP screening is that clinical diagnosis in ROP is subjective with high rates of interobserver variability, and there is inconsistency to real-world treatment differences. The increasing use of fundus photography for recording ROP and in telemedicine initiatives has paved the way for the adoption of artificial intelligence in ROP management.

Artificial intelligence (AI) is a machine algorithm designed to mimick human problem-solving skill. The foundation of artificial intelligence dates back in 1950 when Alan Turing in his paper "Computing machinery and intelligence".[9] Currently, AI is widely used in medicine especially in aiding identification, classification, and diagnosis of various diseases. AI is already developed to aid early diagnosis for diabetic retinopathy,[10] highlighting the potential an AI for retinopathy of prematurity. Increase used of fundus photography in telemedicine ROP screening programs has facilitated the implementation an AI model for diagnosis. AI model has an advantage over human in ROP screening program especially because

computers are not susceptible to fatigue and bias that may affect assessment result, it is a low risk examination.[11] In healthcare economics, AI has shown to reduced overall diagnosis burden of a healthcare by improving diagnostic accuracy, enables early detection with minimal device, and preventing overdiagnosis and overtreatment.[12] Integrating wide-field imaging and automated diagnosis within a teleophthalmology system offers a potential solution to these problems. This approach could facilitate quick screening and prioritization of infants, even in areas with limited resources. Given the prevalence of ROP and the increasing demand for efficient screening solutions, this systematic review aims to update the current development of AI technologies for ROP diagnosis and screening, considering the appropriate AI types that align with the specific needs and workloads of ROP screening programs.

Artificial Intelligence (AI) has emerged as a promising tool in addressing challenges of timely and accurate diagnosis of retinopathy of prematurity (ROP), particularly in resource-limited settings where access to trained specialists may be limited. Its ability to process large volumes of retinal images rapidly and consistently offers potential to improve screening coverage and reduce missed diagnoses. However, despite growing interest, current AI models vary significantly in design, dataset diversity, and validation methods. This systematic review aims to critically evaluate the latest AI models developed for ROP screening, highlight their diagnostic performance, and identify existing limitations in their clinical validation and generalizability.

The primary outcome of this review is the diagnostic performance of artificial intelligence (AI) models in screening for retinopathy of prematurity (ROP). Key performance indicators include the area under the receiver operating characteristic curve (AUC), sensitivity, and specificity. These metrics are critical for evaluating how well AI models can distinguish between diseased and non-diseased cases. High sensitivity is particularly important in the context of ROP screening to minimize the risk of missing sight-threatening cases, while high specificity reduces false positives that may lead to unnecessary referrals or anxiety. Secondary outcomes include additional diagnostic metrics such as inter-rater agreement (to assess consistency between AI and human graders), negative predictive value (NPV), positive predictive value (PPV), and overall diagnostic accuracy. These outcomes reflect how AI tools might perform in real-world clinical settings, particularly in varying disease prevalence and image quality conditions, and help determine the reliability and applicability of AI-assisted screening in different healthcare contexts.

This study is a systematic review study conducted by systematically searching relevant studies through several online database which includes Cochrane, MEDLINE, and ScienceDirect. The search was conducted in 20th April 2024. The PICO of this study is defined as follows: premature infants (Patients) diagnosed by artificial intelligence (AI) models (Intervention) compared with standardized diagnostic methods performed by humans (Comparison), with the outcome being diagnostic performance measured by AUC, sensitivity, and specificity (Outcome). The keywords used in each database is presented in table 1.

Table 1. Search terms in each database

Database	Keywords	Entries found
Cochrane	ID #1 ("artificial intelligence"):ti,ab,kw OR ("deep learning"):ti,ab,kw OR ("machine learning"):ti,ab,kw #2 ("retinopathy of prematurity"):ti,ab,kw OR (ROP):ti,ab,kw #3 ("diagnosis"):ti,ab,kw OR (prediction):ti,ab,kw OR ("sensitivity analysis"):ti,ab,kw OR ("specificity"):ti,ab,kw OR ("area under the curve"):ti,ab,kw #4 #1 AND #2 AND #3	3
MEDLINE	(((((("artificial intelligence"[All Fields]) OR ("machine learning"[All Fields])) OR ("ai"[All Fields])) OR ("convolutional neural network"[All Fields])) AND (((((("diagnosis"[All Fields]) OR ("prediction"[All Fields])) OR ("sensitivity"[All Fields])) OR ("area under the curve"[All Fields])) OR ("screening"[All Fields])) OR ("specificity"[All Fields]))) AND (("retinopathy of prematurity"[All Fields]) OR ("rop"[All Fields]))	64
ScienceDirect	("Artificial intelligence" OR "Machine learning" OR "Deep learning") AND ("Retinopathy of prematurity") AND (Diagnosis OR prediction OR sensitivity OR area under the curve)	518

This study was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines. This systematic review will include studies assessing the capabilities of an AI model to predict and diagnosing retinopathy of prematurity using several relevant clinical parameters. The study must include validation test using the prespecified golden standard and presented the relative capability of the algorithm in detecting retinopathy of prematurity relative to the golden standard. The inclusion criteria for this study were cross-sectional analytical diagnostic or cohort studies that compared the diagnostic capabilities of an AI model for retinopathy of prematurity against a gold standard examination, provided a clear description of both the gold standard and AI model used, and reported the primary outcomes of interest, while exclusion criteria comprised studies without full-text availability, non-English studies, studies limited to AI model generation, and publications in the form of case reports, case series, case-control studies, reviews, editorials, or commentaries.

Risk of bias was assessed using the QUADAS-2 tool, which is designed to evaluate the quality of primary diagnostic accuracy studies. Each study was independently assessed across four domains: (1) patient selection, (2) index test, (3) reference standard, and (4) flow and timing. For each domain, we evaluated the risk of bias and applicability concerns using the signaling questions provided in the QUADAS-2 framework. Discrepancies between reviewers were resolved through discussion and consensus.

In addition to risk of bias assessment, the Oxford Centre for Evidence-Based Medicine (CEBM) 2011 Levels of Evidence were used to classify the overall strength of the included studies (Table 2).[13] A critical appraisal of diagnostic accuracy was also performed using the CEBM checklist to support our interpretation of each study's methodological rigor.

Table 2. Oxford Center of Evidence-Based Medicine 2011 Levels of Evidence.

LOE	Studies
I	Systematic reviews (with homogeneity) of RCT
II	RCT or observational studies with dramatic effect
III	Non-randomized controlled cohort / follow-up studies
IV	Case series, case control, or historically controlled studies
V	Mechanism-based reasoning

LOE: level of evidence

We extracted the information from each study that fulfilled the inclusion and exclusion criteria. Data regarding the author's name, year of publication, study design, type of artificial intelligence used, the training data used, and the outcomes

RESULTS

Figure 1 in this systematic review represents a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart, which details the study selection process. The flowchart begins with the identification phase, where a total of 608 studies were found across several databases including Cochrane, PubMed, and ScienceDirect. A total of 560 studies were excluded during the screening phase. In the eligibility phase, the full texts of these 16 studies were examined in detail. During this process, 4 studies were excluded with two studies excluded because of its literature review design and two studies because it lacks AI model validation test. Critical Appraisal for each included studies is presented in table 4 below.

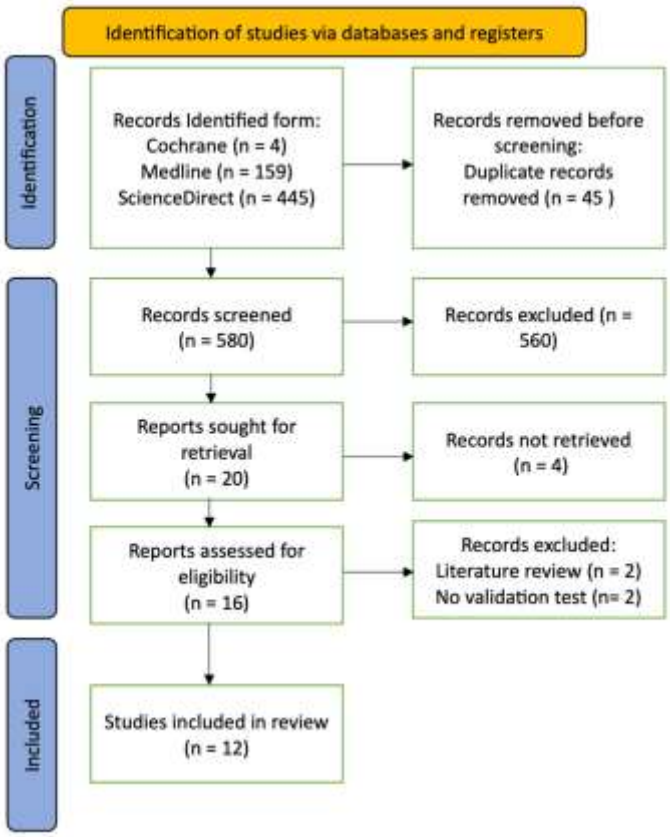


Figure 1. PRISMA Flowchart

Table 4. Critical appraisal of included studies

Study ID	Validity				Applicability
	Was the diagnostic test evaluated in a Representative spectrum of patients (like those in whom it would be used in practice)	Was the reference standard applied regardless of the index test result?	Was there an independent, blind comparison between the index test and an appropriate reference ('gold') standard of diagnosis?		Were the methods for performing the test described in sufficient detail to permit replication?
			Answer	Details	
Brown (2018).[17]	Yes	Yes	Yes	Consensus of image-based grading by three experts	Yes
Greenwald (2020).[18]	Yes	Yes	Yes	Graded by an ophthalmologist using ICROP	Yes
Campbell (2021).[20]	Yes	Yes	Yes	Three clinicians using telemedicine	Yes
Chen (2021).[21]	Yes	Yes	Yes	Consensus diagnosis by three expert graders	Yes
Campbell (2022).[22]	Yes	Yes	Yes	Determined by 34 ROP experts.	Yes
Cole (2022).[23]	Yes	Yes	Yes	Manual diagnosis using ICROP	Yes
Coyner (2022).[24]	Yes	Yes	Yes	Manual diagnosis using ICROP	Yes
Li (2022).[25]	Yes	Yes	Yes	Consensus diagnosis from three ROP experts.	Yes
Bai (2023).[26]	Yes	Yes	Yes	five expert pediatric ophthalmologists	Yes
Liu (2023).[27]	Yes	Yes	Yes	Determined by senior ophthalmologists	Yes

Study ID	Validity				Applicability
	Was the diagnostic test evaluated in a Representative spectrum of patients (like those in whom it would be used in practice)	Was the reference standard applied regardless of the index test result?	Was there an independent, blind comparison between the index test and an appropriate reference ('gold') standard of diagnosis?		Were the methods for performing the test described in sufficient detail to permit replication?
			Answer	Details	
Rao (2023).[28]	Yes	Yes	Yes	Grading by trained ROP graders	Yes
Siegfried (2023).[29]	Yes	Yes	Yes	majority vote of three senior pediatric ophthalmologists.	Yes

The study characteristics table (Table 5) provides a detailed summary of the studies included in this systematic review on the use of artificial intelligence (AI) for diagnosing retinopathy of prematurity (ROP). It encompasses a wide range of study designs, geographical locations, and AI models, offering a comprehensive overview of the current state of research in this area. Overall, these studies illustrate the global effort in utilizing AI for ROP diagnosis, employing various AI models and training datasets to improve diagnostic accuracy and early detection in premature infants. The diversity in study designs, populations, and AI technologies highlights the extensive research dedicated to enhancing the screening and management of ROP through artificial intelligence.

Table 5. Study Characteristics

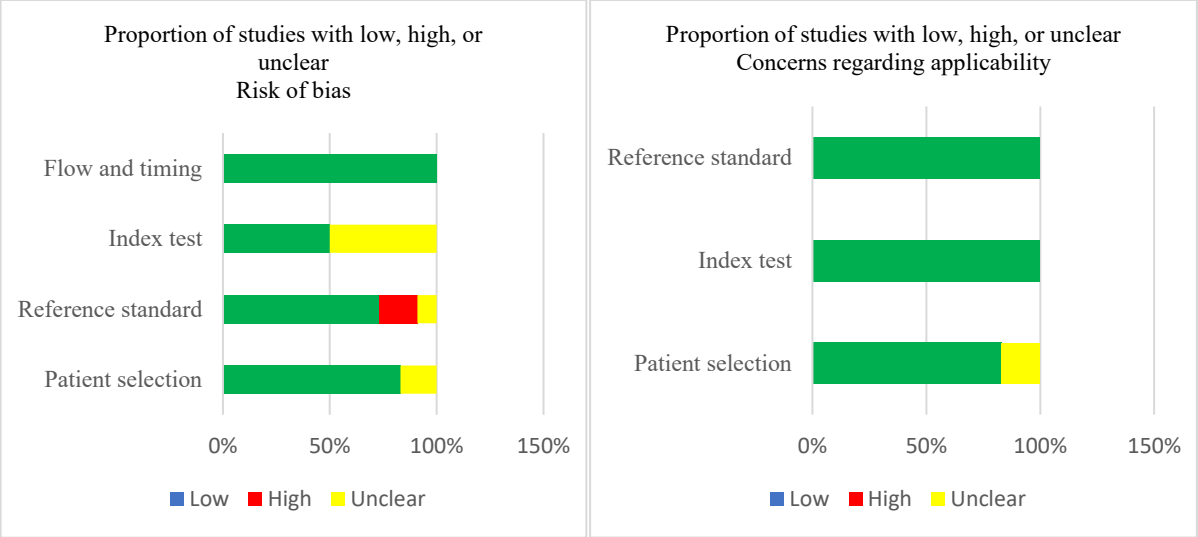
Study ID	Study Design	Country	Mean PMA/GA Age (weeks)	ROP type	Number of samples	Training data source	Model name
Brown (2018).[17]	Cross-sectional	USA	N/A	No plus, pre-plus, and plus disease ROP	5511 images	Retinal image	i-ROP DL
Greenwald (2020).[18]	Cross-sectional	USA	29.2 ± 2.1	Type 1 & Type 2	79 without ROP 2 with ROP	Retinal images	i-ROP DL
Campbell (2021).[20]	Cross-sectional	India	31.6 ± 4	No plus, pre-plus, and plus disease ROP	4175 images from 1253 eyes	Retinal images	i-ROP DL
Chen (2021).[21]	Cross-sectional	North America and Nepal	North America: 26.6 ± 2.2 Nepal: 32.6 ± 2.8	Stage 1-3 ROP	10894 images	Retinal images	ResNet-152

Study ID	Study Design	Country	Mean PMA/GA Age (weeks)	ROP type	Number of samples	Training data source	Model name
Campbell (2022).[22]	Cross-sectional	USA	Unspecified	Stage 1-5 No plus, pre-plus, and plus ROP	Unspecified	Retinal images	i-ROP DL
Cole (2022).[23]	Cross-sectional	Nepal and Mongolia	Nepal: 33.3 ± 2.5 Mongolia: 30.4 ± 2.1	No plus, pre-plus, and plus ROP	Nepal: 391 eyes Mongolia: 467 eyes	Retinal images	i-ROP DL
Coyner (2022).[24]	Cross-sectional	India, Nepal, Mongolia	Not treated: 33.5 ± 2.8 Treated: 29.7 ± 2.2	Unspecified	Not treated: 3633 patients Treated: 127 patients	Retinal images	No name
Li (2022).[25]	Cross-sectional	China	31.31 ± 5.42	Stage 1-3 ROP	Training set: 14,626 images Test set: 3680 images Comparison set: 521 images	Retinal images	Dense Net
Bai (2023).[26]	Retrospective Cohort	Australia	27.74 ± 2.82	ROP	8052 images	Retinal images	ROP.AI
Liu (2023).[27]	Retrospective Cohort	China	N/A	Treatment indicated ROP	24,495 images from 1075 eyes	Retinal image	ResNet-18 and DenseNet-121
Rao (2023).[28]	Cross-sectional	India	N/A	ROP	7,489 images	Retinal image	EfficientNet-B0
Siegfried (2023).[29]	Retrospective Cohort	UK	Less than 32 weeks old or birthweight less than 1501 gram	No plus, pre-plus, and plus ROP	Training set: 6141 images Test set: 200 images	Retinal image	Bespoke and CFDL model

UK: United Kingdom; PMA: post-menstrual age; GA: gestational age; ROP: retinopathy of prematurity

The risk of bias assessment for the diagnostic studies included in this systematic review was conducted using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) (Table 6). QUADAS-2 is designed with four main domains, each evaluated for risk of bias and relevance to the research question. There are four points evaluated for risk of bias are patients selection, index test, references standard, and flow timing. To aid in assessing these aspects, each domain includes a set of signalling questions. There remained a potential risk of bias due to the following factors: (1)inappropriate exclusions during patient selection; (2) not all subjects were included in the analysis; (3) qualifications of the examiners were not specified

Table 6. Risk of bias assessment of cohort studies using QUADAS-2



Brown (2018), Greenwald (2020), Campbell (2021 and 2022), and Cole (2022) used i-ROP Deep Learning (DL) model for ROP diagnosis (Table 7). The earliest study that uses this model was a study by Brown (2018). The model utilized two primary neural network architectures: a vessel segmentation network and a classification network. The vessel segmentation network was designed using the U-Net architecture, which is highly specialized for biomedical image segmentation.[17] The study by Chen (2021) aimed to develop a deep learning model for diagnosing ROP, specifically focusing on identifying stages 1-3 in retinal images of preterm infants.[21] Coyner (2022) study focused on developing and validating a deep learning-based model to screen for ROP in infants from low- and middle-income countries (LMICs).[24] Li (2022) developed an automatic deep convolutional neural network (DCNN)-based system for early diagnosis and quantitative analysis of ROP. Using 18,827 retinal images from preterm infants, two modified Retina U-Nets were employed to segment blood vessels and demarcation lines.[25] Bai (2023) aimed to test the ROP.AI model in Australian population. ROP.AI was developed using retinal images collected from a single center in New Zealand. The algorithm employs convolutional neural networks (CNNs) to analyze retinal images and detect features indicative of plus disease.[26] Liu (2023) study aimed to develop an AI system for identifying disease status and recommending treatment modalities for retinopathy of prematurity (ROP). The AI system's tasks included ROP identification, severe ROP identification, and treatment modality identification (retinal laser photocoagulation or intravitreal injections).[27] Rao (2023) study aimed to develop and validate an AI-based screening tool for detecting ROP in South Indian infants. They employed convolutional neural networks (CNNs), specifically the EfficientNet-B0 architecture, to train a deep learning algorithm

capable of binary classification (ROP present vs. ROP absent).[28] Siegfried (2023) aimed to develop and validate deep learning models for detecting plus disease in ROP. Two types of models were developed: bespoke and code-free deep learning (CFDL) models. Lastly, the CFDL model was developed using Google Cloud AutoML Vision API, which does not require advanced coding skills, making it accessible for use in low-resource settings.[29]

Table 7 provides an in-depth analysis of the diagnostic performance of various AI models used in detecting retinopathy of prematurity (ROP) across multiple studies. This detailed examination

Table 7. Diagnostic performace of AI model

Study ID	AI model	Detection	Se (%)	Sp (%)	AUC	Additional metrics	Comments
Brown (2018).[17]	i-ROP DL	Plus disease	93.0	94.0	0.94	Inter-rater agreement 0.92	i-ROP DL shows high diagnostic accuracy with higher agreement than 6 out of 8 experts
		Pre-plus disease	100	94.0	0.98		
Greenwald (2020).[18]	i-ROP DL	Referral-requiring ROP (Type 1 and 2 ROP)	100	90.0	0.99	-	Severity score above 3 is highly predictive for ROP early detection
Campbell (2021).[20]	i-ROP DL	Treatment-requiring ROP (with plus disease)	100	78.0	0.98	PPV 12%	AI can be effectively integrated into telemedicine programs to enhance screening efficiency and monitor disease
Chen (2021).[21]	ResNet-152 Nepal data set	Stage 1-3 ROP	98.0	96.0	0.98		The study highlights the domain shift, significant drops of AI model performance when tested in different population/different camera
	ResNet-152 North America data set		82.0	99.0	0.99		
Campbell (2022).[22]	i-ROP DL	ROP	-	-	-	inter-rater agreement 0.67 Correlation coefficient 0.88 for disease severity	The deep learning-derived vascular severity score showed strong consistency with expert classifications
Cole (2022).[23]	i-ROP DL Nepal data set	Plus ROP	75.0	64.5	0.99		This study demonstrated that the i-ROP DL algorithm performed well across different camera systems and populations
	i-ROP DL Mongolia data set	Plus ROP	89.3	54.3	0.97		
Coyner (2022).[24]	No name India data set	ROP	100	63.3	-		Varying specificity indicates room for improvement to reduce false positives
	No name Nepal data set		100	77.8	-		
	No name Mongolia data set		100	45.8	-		
Li (2022).[25]	Dense Net	Normal images	95.9	96.4	0.96	Inter-rater agreement 0.94	The system's ability to quantitatively analyze features such as the width of demarcation lines and vascular bifurcation ratios provides an objective basis for diagnosis
		Stage I ROP	90.2	97.7	0.93		
		Stage II ROP	92.8	98.7	0.99		
		Stage III ROP	91.8	99.3	0.99		
Bai (2023).[26]	ROP.AI	Plus ROP	84.0	43.0	0.75	NPV 96%	The relatively low specificity indicates a higher rate of false positives. Misclassifications often occurred in images with darker fundus or slight blurring.
Liu (2023).[27]	ResNet-18 and DenseNet-121	ROP	85.9	91.7	95.3	Accuracy 88.5%	The AI system outperformed experienced ophthalmologists in accuracy, especially in determining the need for treatment and selecting the appropriate treatment modality for ROP
		Severe ROP	98.0	52.4	91.3	Accuracy 84.7%	
		Treatment modality identification	70.6	94.1	93.6	Accuracy 86.3%	
Rao (2023).[28]	EfficientNet-B0	ROP	91.5	91.2	0.97	PPV 81.7% NPV 96.14	The false negatives in the test set were mainly from Stage 1 and Stage 2 ROP, which are harder to detect due to subtle features. Ensuring high-quality images is crucial for the model's performance.
Siegfried (2023).[29]	Bespoke	Healthy	-	-	0.98	Inter-rater agreement 0.77	The study found high inter-observer variability, especially among less experienced clinicians. This variability underscores the challenge in establishing a consistent reference standard for training and validating AI models
		Pre-plus ROP	-	-	0.93		
		Plus ROP	-	-	0.97		
	CFDL	Healthy	-	-	0.99	Inter-rater agreement 0.53	
		Pre-plus ROP	-	-	0.93		
		Plus ROP	-	-	0.98		

DISCUSSION

This systematic review aimed to evaluate the diagnostic performance of various AI models in detecting ROP. Across multiple studies, AI models consistently demonstrated high sensitivity, highlighting their strong potential for early detection of ROP. Prioritizing sensitivity and comparing results to the accuracy of the reference standard within the same population aligns with the standard approach in diagnostic studies. Sensitivity values of 100% reported in Brown (2018)[17] and Greenwald (2020)[18] confirm the highest sensitivity potential of AI models, consistent with earlier findings on AI use in ophthalmology. In the broader field of eye disease, AI is mainly used for diagnosis, including in glaucoma and diabetic retinopathy.

Performance, however, varied across studies. While the i-ROP DL model showed excellent sensitivity in several reports, specificity ranged considerably. The ResNet-152 model demonstrated a marked drop in sensitivity when applied to a different population, as seen in Chen (2021)[21]. Such variability indicates that although AI models are effective at identifying true positives, calibration is needed to reduce false positives. High AUC values reported in Li (2022)[25] and Siegfried (2023)[29] indicate generally good discriminatory ability, with accuracy categories ranging from excellent (90–100%) to very poor (50–60%). Nonetheless, lower PPV in some studies, such as Campbell (2021)[20], suggests that high false-positive rates remain a concern, particularly in screening contexts where low specificity could lead to unnecessary diagnostic procedures.

While AI holds promise for ROP screening, several barriers remain. Image quality is a critical factor influencing diagnostic accuracy, especially in low-resource settings where imaging equipment may be suboptimal. Training personnel to produce high-quality images before implementation is strongly recommended. Variability in model performance across clinical environments reflects differences in patient populations, imaging systems, and disease prevalence, making standardized performance difficult to achieve.

Furthermore, many included studies were retrospective, which may limit real-world applicability. The lack of standardized performance reporting across studies also complicates direct model comparisons. Although AI is often assumed to be cost-efficient for large-scale screening, particularly in developing countries, none of the included studies explicitly assessed cost-effectiveness.

This review identified clear evidence of domain shift, where models trained on one dataset or population perform less accurately when applied to another. The drop in sensitivity for the ResNet-152 model in Chen (2021)[21] exemplifies this problem. Domain shift occurs when external samples differ in features from the original training dataset, leading to reduced performance. Such findings underscore the need for domain adaptation, which may involve fine-tuning models with local data or training on large, diverse, multi-center datasets. Differences in demographic composition, disease spectrum, and imaging hardware across studies limit generalizability, reinforcing the importance of validating AI models in the specific populations and clinical settings where they will be used.

When effectively deployed, AI-assisted ROP screening offers substantial clinical benefits. High sensitivity ensures most true cases are detected, enabling timely interventions such as anti-VEGF injection, cryotherapy, or surgery. AI applications in telemedicine, demonstrated by Campbell (2021)[20] and Greenwald (2020)[18], can improve accessibility in remote and low-resource areas. AI can also standardize grading, particularly in differentiating plus from non-plus disease, reducing inter-observer

variability and ensuring consistent diagnostic thresholds. Automating initial screenings can reduce ophthalmologists' workload, allowing specialists to focus on complex cases. However, these benefits depend on maintaining high image quality, ensuring robust validation in target populations, and achieving acceptable specificity to prevent unnecessary follow-up procedures.

Future research should aim to improve AI specificity without compromising sensitivity, thereby reducing false positives. Large, prospective, multi-center trials involving diverse demographics and multiple imaging systems will be essential for validating performance and improving robustness. Advancing domain adaptation techniques will help mitigate population and equipment differences. Integrating AI outputs with other diagnostic tools and clinical data may enable comprehensive ROP assessments, prediction of infants at highest risk for severe disease, and personalized treatment planning. Further research should also explicitly evaluate cost-effectiveness, particularly in low- and middle-income countries. Standardizing performance metrics, maintaining high image quality, and providing healthcare worker training in AI-assisted workflows will be crucial to maximize clinical impact. By addressing these factors, AI can evolve from a promising diagnostic tool to an integral component of ROP management.highlights the effectiveness of these AI models in identifying different stages and severities of ROP, emphasizing their sensitivity, specificity, area under the curve (AUC), and additional diagnostic metrics.

Brown (2018) and Greenwald (2020) utilized the i-ROP Deep Learning (DL) model, achieving high sensitivity and specificity for plus and pre-plus disease ROP. Greenwald reported perfect sensitivity (100%) and high specificity (90%) with an AUC of 0.99 for referral-requiring ROP, underscoring the model's potential for early detection. Campbell (2021) also employed the i-ROP DL model, focusing on treatment-requiring ROP, with 100% sensitivity and 78% specificity (AUC 0.98). This study highlighted the model's effectiveness in telemedicine, despite a lower PPV of 12%. Chen (2021) used the ResNet-152 model on datasets from Nepal and North America. For Nepal, the model achieved 98% sensitivity and 96% specificity (AUC 0.98). Campbell (2022), continuing with i-ROP DL, reported strong consistency with expert classifications but did not provide specific metrics, emphasizing high inter-rater agreement. Cole (2022) evaluated i-ROP DL in Nepal and Mongolia. Coyner (2022) developed a model tested in India, Nepal, and Mongolia, achieving 100% sensitivity but varying specificity (63.3% for India, 77.8% for Nepal, and 45.8% for Mongolia), suggesting high sensitivity but the need for improved specificity. Rao (2023) used EfficientNet-B0, achieving 91.5% sensitivity and 91.2% specificity (AUC 0.97).

Overall, the AI models demonstrate high sensitivity across various studies, indicating their strong potential for early detection of ROP. However, variability in specificity and other metrics such as PPV and NPV suggests that while these models are effective in identifying true positives, there is a need for further refinement to reduce false positives. This is particularly important in clinical settings to avoid unnecessary treatments and interventions. The consistent high performance of models like i-ROP DL and ResNet across different studies and populations underscores their reliability. The integration of AI models in telemedicine and clinical workflows, as suggested by studies like Campbell (2021) and Greenwald (2020), can enhance screening efficiency and improve the management of ROP. The use of AI in low-resource settings, as explored by Coyner (2022) and Rao (2023), demonstrates its potential to bridge gaps in healthcare access and quality

CONCLUSION

This systematic review, conducted to evaluate the diagnostic performance of AI models for ROP detection, found that most models achieve consistently high sensitivity, supporting their potential as effective early screening tools. However, marked variability in specificity and positive predictive value across studies highlights the need to refine algorithms to reduce false positives and improve clinical applicability. Evidence of domain shift underscores that AI models must be validated and, where necessary, adapted to the target population and imaging systems before deployment. Clinically, these findings suggest that AI could expand screening coverage, standardize grading, and facilitate telemedicine-based ROP programs, particularly in low-resource settings, provided image quality standards and workflow integration are ensured. For research, the results point to the need for prospective, multi-center studies that include diverse demographics, standardized performance metrics, and cost-effectiveness analyses to confirm generalizability and guide large-scale implementation.

,

REFERENCES

1. Brown AC, Nwanyanwu K. Retinopathy of Prematurity. In Treasure Island (FL); 2024.
2. Good W V. Final results of the Early Treatment for Retinopathy of Prematurity (ETROP) randomized trial. Trans Am Ophthalmol Soc. 2004;102:233–50.
3. Siswanto JE, Bos AF, Dijk PH, Rohsiswatmo R, Irawan G, Sulistijono E, et al. Multicentre survey of retinopathy of prematurity in Indonesia. BMJ Paediatr Open. 2021;5:e000761.
4. Blencowe H, Lawn JE, Vazquez T, Fielder A, Gilbert C. Preterm-associated visual impairment and estimates of retinopathy of prematurity at regional and global levels for 2010. Pediatr Res. 2013 Dec;74 Suppl 1(Suppl 1):35–49.
5. Vos T, Flaxman AD, Naghavi M, Lozano R, Michaud C, Ezzati M, et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet (London, England). 2012 Dec;380(9859):2163–96.
6. Fierson WM. Screening Examination of Premature Infants for Retinopathy of Prematurity. Pediatrics. 2018 Dec;142(6).
7. Dewi R, Tuhusula R, Rohsiswatmo R. Skrining Retinopathy of Prematuritydi Rumah Sakit dengan Fasilitas Terbatas. Sari Pediatr. 2016;14(3):185.
8. Scruggs BA, Chan RVP, Kalpathy-Cramer J, Chiang MF, Campbell JP. Artificial Intelligence in Retinopathy of Prematurity Diagnosis. Transl Vis Sci Technol. 2020 Feb;9(2):5.
9. Rockwell A. The History of Artificial Intelligence [Internet]. Harvard University. 2017. Available from: https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/
10. Padhy SK, Takkar B, Chawla R, Kumar A. Artificial intelligence in diabetic retinopathy: A natural step to the future. Indian J Ophthalmol. 2019 Jul;67(7):1004–9.
11. Gensure RH, Chiang MF, Campbell JP. Artificial intelligence for retinopathy of prematurity. Curr Opin Ophthalmol. 2020 Sep;31(5):312–7.
12. Khanna NN, Maindarkar MA, Viswanathan V, Fernandes JFE, Paul S, Bhagawati M, et al. Economics of Artificial Intelligence in Healthcare: Diagnosis vs. Treatment. Healthc (Basel, Switzerland). 2022 Dec;10(12).

13. Oxford Centre for Evidence-Based Medicine: Levels of Evidence (March 2009) [Internet]. University of Oxford. 2021 [cited 2021 Sep 20]. Available from: https://www.cebm.ox.ac.uk/resources/levels-of-evidence/oxford-centre-for-evidence-based-medicine-levels-of-evidence-march-2009

14. ExploreAI. Artificial Intelligence (AI) [Internet]. 2023. Available from: https://exploreai.org/p/ai-definition

15. IBM. What is deep learning [Internet]. Available from: https://www.ibm.com/topics/deep-learning

16. Trevethan R. Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. Front Public Heal. 2017;5(November):1–7.

17. 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):803–10.

18. Greenwald MF, Danford ID, Shahrawat M, Ostmo S, Brown J, Kalpathy-Cramer J, et al. Evaluation of artificial intelligence-based telemedicine screening for retinopathy of prematurity. J AAPOS. 2020;24(3):160–2.

19. Lepore D, Ji MH, Pagliara MM, Lenkowicz J, Capocchiano ND, Tagliaferri L, et al. Convolutional neural network based on fluorescein angiography images for retinopathy of prematurity management. Transl Vis Sci Technol. 2020;9(2):1–8.

20. Campbell JP, Singh P, Redd TK, Brown JM, Shah PK, Subramanian P, et al. Applications of artificial intelligence for retinopathy of prematurity screening. Pediatrics. 2021;147(3).

21. Chen JS, Coyner AS, Ostmo S, Sonmez K, Bajimaya S, Pradhan E, et al. Deep Learning for the Diagnosis of Stage in Retinopathy of Prematurity: Accuracy and Generalizability across Populations and Cameras. Ophthalmol Retin. 2021 Oct;5(10):1027–35.

22. Campbell JP, Chiang MF, Chen JS, Moshfeghi DM, Nudleman E, Ruambivoonsuk P, et al. Artificial Intelligence for Retinopathy of Prematurity: Validation of a Vascular Severity Scale against International Expert Diagnosis. Ophthalmology. 2022 Jul;129(7):e69–76.

23. 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 [Internet]. 2022;2(4):100165. Available from: https://doi.org/10.1016/j.xops.2022.100165

24. 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):791–8.

25. 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):1–12.

26. 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):1–9.

27. Liu Y, Du Y, Wang X, Zhao X, Zhang S, Yu Z, et al. An Artificial Intelligence System for Screening and Recommending the Treatment Modalities for Retinopathy of Prematurity. Asia-Pacific J Ophthalmol [Internet]. 2023;12(5):468–76. Available from: http://dx.doi.org/10.1097/APO.0000000000000638

28. Rao DP, Savoy FM, Tan JZE, Fung BPE, 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 [Internet]. 2023;11(September):1–11. Available from: https://doi.org/10.3389/fped.2023.1197237

29. 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 Heal [Internet]. 2023;5(6):e340–9. Available from: http://dx.doi.org/10.1016/S2589-7500(23)00050-X

30. Kiran Yenice E, Kara C, Yenice M, Erdas CB. Retinopathy of Prematurity in Late Preterm Twins with a Birth Weight Discordance: Can it be Predicted by Artificial Intelligence? Beyoglu Eye J. 2023;8(4):287–92.

31. Liu H, Li L, Wormstone IM, Qiao C, Zhang C, Liu P, et al. Development and Validation of a Deep Learning System to Detect Glaucomatous Optic Neuropathy Using Fundus Photographs. JAMA Ophthalmol. 2019;137(12):1353–60.

32. Lim JI, Regillo CD, Sadda SR, Ipp E, Bhaskaranand M, Ramachandra C, et al. Artificial Intelligence Detection of Diabetic Retinopathy: Subgroup Comparison of the EyeArt System with Ophthalmologists' Dilated Examinations. Ophthalmol Sci. 2023 Mar;3(1):100228..

This work licensed under Creative Commons Attribution