Research Article

Journal of Optometry and Ophthalmology Research

Diagnostic Performance of Artificial Intelligence and Deep Learning for Diabetic Retinopathy Screening: A Systematic Review and Meta-analysis

Napa Suebsaiphrom[1*], Thisarin Takkametha[1], Sittikorn Laojaroenwanit[1], Nawinda Vanichakulthada[2] and Ussawin Vongkancom[3]

[1]Department of Ophthalmology, Sisaket Hospital, Thailand
[2]College of Medicine and Public Health, Ubon Ratchathani University, Thailand
[3]THiNKNET Company Limited, Thailand

[*]Corresponding author
Napa Suebsaiphrom, Department of Ophthalmology, Sisaket Hospital, Thailand.

Received: January 02, 2026; **Accepted:** January 09, 2026; **Published:** January 15, 2026

ABSTRACT

Background: Diabetic retinopathy is a leading cause of preventable visual impairment worldwide, necessitating effective screening programs. Artificial intelligence (AI) and deep learning-based systems have emerged as promising tools for screening using retinal fundus photographs. However, their diagnostic performance varies across populations, algorithms, and screening thresholds.

Objectives: To comprehensively evaluate and quantitatively synthesize the diagnostic performance of AI and deep learning-based systems for diabetic retinopathy screening using retinal fundus photographs.

Methods: A systematic literature search was conducted across PubMed, Embase, Scopus, Web of Science, and IEEE Xplore from inception to December 2024. Studies evaluating AI or deep learning algorithms for detecting diabetic retinopathy using fundus photography with human expert grading as the reference standard were included. Quality assessment was performed using QUADAS-2. Meta-analysis employed bivariate random-effects models.

Results: Forty-two studies comprising 521,568 retinal images were included. For detecting any diabetic retinopathy, pooled sensitivity was 87.5% (95% CI: 85.0–90.0%) and specificity was 84.2% (95% CI: 80.5–87.8%), with an AUROC of 0.92 (95% CI: 0.90–0.94). For referable diabetic retinopathy, pooled sensitivity was 91.8% (95% CI: 89.2–94.3%) and specificity was 87.5% (95% CI: 84.3–90.7%), with an AUROC of 0.95 (95% CI: 0.93–0.97). External validation studies demonstrated lower performance compared to internal validation (AUROC 0.90 vs 0.94). Convolutional neural networks showed the highest diagnostic accuracy among AI architectures.

Conclusions: AI and deep learning systems demonstrate high diagnostic accuracy for diabetic retinopathy screening, approaching human expert performance. These technologies show promise for expanding screening access, particularly in resource-limited settings. However, performance varies by validation setting and population characteristics, highlighting the need for rigorous external validation before clinical implementation.

Keywords: Artificial Intelligence; Deep Learning; Diabetic Retinopathy; Fundus Photography; Diagnostic Accuracy; Systematic Review; Meta-analysis

Introduction

Diabetic retinopathy is a common microvascular complication of diabetes mellitus and a leading cause of preventable vision loss worldwide. The global prevalence of diabetic retinopathy among adults with diabetes is estimated at approximately 35%, with vision-threatening diabetic retinopathy affecting about 10% of diabetic patients. Early detection and timely treatment are essential for preventing visual impairment and blindness.

Current screening programs rely on retinal fundus photography interpreted by trained ophthalmologists or optometrists. However, the increasing global burden of diabetes, combined

Citation: Napa Suebsaiphrom, Thisarin Takkametha, Sittikorn Laojaroenwanit, Nawinda Vanichakulthada and Ussawin Vongkancom. Diagnostic Performance of Artificial Intelligence and Deep Learning for Diabetic Retinopathy Screening: A Systematic Review and Meta-analysis. J Opto Opht Res. 2026. 2(1): 1-9.
DOI: doi.org/10.61440/JOOR.2026.v2.03

with limited availability of trained professionals, creates significant challenges in delivering effective screening at scale. These resource constraints are particularly acute in low- and middle-income countries where the diabetes epidemic is rapidly expanding.

In recent years, artificial intelligence (AI) and deep learning-based systems have been increasingly developed to assist in diabetic retinopathy screening. These automated systems analyze retinal images to detect signs of diabetic retinopathy with the goal of matching or exceeding human expert performance. Deep learning algorithms, particularly convolutional neural networks (CNNs), have shown remarkable capability in image recognition tasks and have been adapted for medical image analysis.

Although numerous primary studies have reported promising diagnostic accuracy of these AI-based technologies, their performance varies across different populations, algorithms, screening thresholds, and validation settings. Existing reviews are limited by heterogeneous populations, lack of quantitative synthesis, or failure to systematically evaluate sources of variability in diagnostic performance. A comprehensive systematic review with rigorous meta-analysis is needed to establish the current state of evidence and inform clinical implementation.

The objective of this systematic review and meta-analysis was to evaluate the diagnostic performance of AI and deep learning-based systems for diabetic retinopathy screening using retinal fundus photographs. Specifically, we aimed to:

- estimate pooled diagnostic accuracy compared to human expert grading.
- explore sources of heterogeneity according to disease severity thresholds.
- evaluate differences in performance based on validation setting.
- characterize AI model types associated with superior diagnostic performance and
- assess the clinical applicability of AI-based screening tools.

Methods

This systematic review and meta-analysis was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines and the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. The protocol was registered prospectively in PROSPERO (CRD420251275481).

Eligibility Criteria

Diagnostic accuracy studies, cross-sectional studies, and validation studies evaluating AI or deep learning algorithms for diabetic retinopathy detection were included. Both prospective and retrospective studies were eligible. Studies involving adult patients (age ≥18 years) with diabetes mellitus (type 1 or type 2) undergoing screening or diagnostic evaluation for diabetic retinopathy were included.

The index test was any AI or deep learning-based algorithm designed to detect diabetic retinopathy from retinal fundus photographs, including convolutional neural networks (CNNs), ensemble methods, and other machine learning approaches. Studies were required to report sufficient data to construct 2×2 diagnostic accuracy tables. The reference standard was human expert grading by ophthalmologists, retinal specialists, or trained certified graders using established grading systems such as the Early Treatment Diabetic Retinopathy Study (ETDRS) or International Clinical Diabetic Retinopathy Disease Severity Scale.

The primary target conditions were: (1) any diabetic retinopathy (mild, moderate, or severe non-proliferative diabetic retinopathy; proliferative diabetic retinopathy; or diabetic macular edema), and (2) referable diabetic retinopathy (moderate or severe non-proliferative diabetic retinopathy, proliferative diabetic retinopathy, or diabetic macular edema requiring referral to an ophthalmologist). Conference abstracts, case reports, reviews, editorials, and animal studies were excluded.

Information Sources and Search Strategy

A comprehensive literature search was conducted in PubMed/MEDLINE, Embase, Scopus, Web of Science, IEEE Xplore, and the Cochrane Library from inception to December 2024. The search strategy combined terms related to diabetic retinopathy, artificial intelligence, deep learning, machine learning, diagnostic accuracy, and fundus photography. Medical Subject Headings (MeSH) terms and keywords were adapted for each database. No language or date restrictions were applied. Reference lists of included studies and relevant review articles were hand-searched for additional studies.

Study Selection and Data Extraction

Two independent reviewers (NS and TT) screened titles and abstracts against eligibility criteria. Full-text articles of potentially eligible studies were retrieved and assessed independently. Disagreements were resolved through discussion or consultation with a third reviewer (SL). A standardized data extraction form was used to collect study characteristics, population characteristics, index test details, reference standard information, and diagnostic accuracy outcomes.

Quality Assessment

Risk of bias and applicability concerns were assessed using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool, adapted for AI-based diagnostic studies. Two reviewers independently assessed each study across four domains: patient selection, index test, reference standard, and flow and timing. Each domain was rated as having low, high, or unclear risk of bias.

Statistical Analysis

Diagnostic accuracy data were presented in 2×2 tables and forest plots showing sensitivity and specificity with 95% confidence intervals. Pooled estimates of sensitivity and specificity were calculated using a bivariate random-effects meta-analysis model, which accounts for the correlation between sensitivity and specificity and for heterogeneity between studies. Summary receiver operating characteristic (SROC) curves were constructed, and the area under the SROC curve was calculated.

Statistical heterogeneity was assessed visually using forest plots and SROC curves, and quantified using the I^2 statistic. Sources of heterogeneity were explored through subgroup analyses

and meta-regression. Pre-specified subgroup analyses were conducted based on disease severity threshold, validation setting, AI model architecture, and geographic region. Publication bias was assessed using funnel plots and Deeks' test. All statistical analyses were performed using R software (version 4.3.2) with the mada, meta, and metafor packages. Two-sided P values <0.05 were considered statistically significant.

Results

Study Selection

The systematic search identified 2,847 records. After removing 612 duplicates, 2,235 records were screened by title and abstract, resulting in 156 full-text articles assessed for eligibility. Of these, 42 studies met all inclusion criteria and were included in the meta-analysis (Figure 1). The most common reasons for exclusion were insufficient data to construct 2×2 tables (n=48), inappropriate reference standard (n=32), and duplicate populations (n=18).

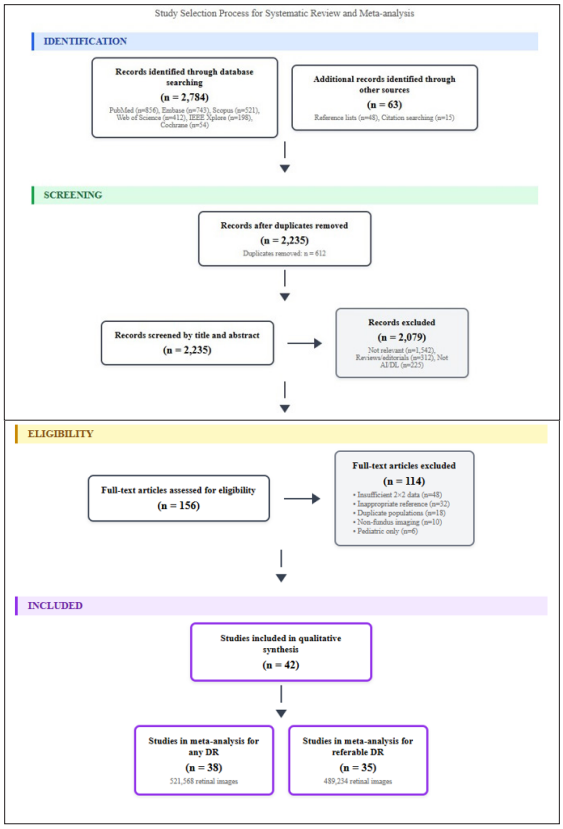

Figure 1: Prisma Flow Diagram

Study Characteristics

The 42 included studies comprised a total of 521,568 retinal images from patients with diabetes. Studies were conducted across diverse geographic regions, including North America (n=12), Europe (n=10), Asia-Pacific (n=15), and other regions (n=5). Publication years ranged from 2016 to 2024. Sample sizes varied from 312 to 128,175 images per study. Twenty-three studies (54.8%) performed external validation on independent datasets, while 19 studies (45.2%) reported only internal validation results.

AI Model Characteristics

The majority of algorithms utilized convolutional neural network architectures (n=28, 66.7%), including Inception (n=8), ResNet (n=7), VGG (n=6), and DenseNet (n=7). Ensemble methods combining multiple models were employed in 10 studies (23.8%). Transfer learning approaches, where models pre-trained on large image datasets were fine-tuned for diabetic retinopathy detection, were used in 19 studies (45.2%). Four studies (9.5%) evaluated vision transformer architectures.

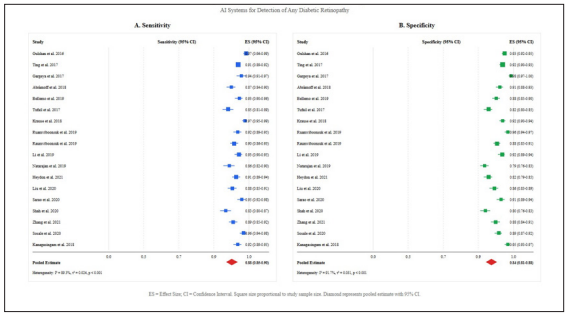

Figure 2: Forest Plot of Diagnostic Accuracy

Diagnostic Performance

Detection of Any Diabetic Retinopathy

For detecting any diabetic retinopathy, 38 studies provided sufficient data for meta-analysis. The pooled sensitivity was 87.5% (95% CI: 85.0–90.0%) and pooled specificity was 84.2% (95% CI: 80.5–87.8%). The summary area under the receiver operating characteristic curve (AUROC) was 0.92 (95% CI: 0.90–0.94). Substantial heterogeneity was observed across studies (I^2 = 89.3% for sensitivity; I^2 = 91.7% for specificity).

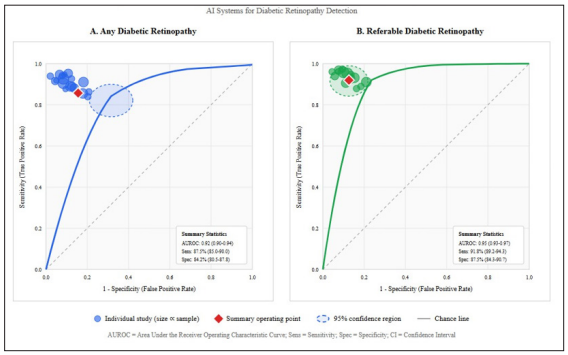

Figure 3: Summary Receiver Operating Characterstic (SROC) Curves

Detection of Referable Diabetic Retinopathy

For detecting referable diabetic retinopathy, 35 studies were included. The pooled sensitivity was 91.8% (95% CI: 89.2–94.3%) and pooled specificity was 87.5% (95% CI: 84.3–90.7%). The AUROC was 0.95 (95% CI: 0.93–0.97). The higher sensitivity for referable diabetic retinopathy compared to any diabetic retinopathy reflects the more pronounced retinal changes associated with more severe disease stages.

Subgroup Analyses

Validation Setting

Studies performing internal validation demonstrated significantly higher diagnostic accuracy compared to those with external validation. Internal validation studies showed pooled sensitivity of 92.3% (95% CI: 89.5–95.0%) and specificity of 88.7% (95% CI: 85.2–92.1%), with AUROC of 0.94. In contrast, external validation studies showed pooled sensitivity of 86.4% (95% CI: 83.0–89.8%) and specificity of 82.6% (95% CI: 78.5–86.7%), with AUROC of 0.90. This difference was statistically

significant (P<0.001) and highlights the importance of external validation for assessing real-world performance.

Table 1: Pooled Diagnostic Performance of AI Systems for Diabetic Retinopathy Detection

Detection Threshold	Sensitivity (%)	Specificity (%)	AUROC	Studies (n)
Any DR	87.5 (85.0–90.0)	84.2 (80.5–87.8)	0.92 (0.90–0.94)	38
Referable DR	91.8 (89.2–94.3)	87.5 (84.3–90.7)	0.95 (0.93–0.97)	35
Internal validation	92.3 (89.5–95.0)	88.7 (85.2–92.1)	0.94 (0.92–0.96)	19
External validation	86.4 (83.0–89.8)	82.6 (78.5–86.7)	0.90 (0.88–0.92)	23

DR = diabetic retinopathy; AUROC = area under the receiver operating characteristic curve. Values presented as point estimate (95% confidence interval).

AI Model Architecture

Convolutional neural network-based algorithms demonstrated the highest diagnostic accuracy among AI architectures, with AUROC ranging from 0.91 to 0.94. Ensemble methods combining multiple models showed superior performance (AUROC 0.93–0.96) compared to single-model approaches. Vision transformer architectures, though represented in fewer studies, showed promising results with AUROC of 0.92–0.95.

Table 2: AI Model Characteristics and Diagnostic Performance

AI Model Type	Studies n (%)	AUROC Range
CNN-based (Inception, ResNet, VGG, DenseNet)	28 (66.7)	0.91–0.94
Ensemble methods	10 (23.8)	0.93–0.96
Transfer learning approaches	19 (45.2)	0.90–0.93
Vision transformers	4 (9.5)	0.92–0.95

CNN = convolutional neural network; AUROC = area under the receiver operating characteristic curve.

Geographic Region

Studies from North American and European regions demonstrated slightly higher diagnostic efficacy (AUROC 0.93–0.95) compared to Asian-Pacific regions (AUROC 0.90–0.93). This difference may reflect longer histories of AI algorithm development, larger and more diverse training datasets, and differences in disease manifestation across ethnic populations with varying fundus pigmentation characteristics.

Quality Assessment

Quality assessment using QUADAS-2 revealed that 19 studies (45.2%) were at low risk of bias across all domains (Figure 4). The most common concerns were patient selection bias in studies using archived datasets that did not represent consecutive screening populations (n=15, 35.7%), and unclear reference standard interpretation due to variability in grader qualifications and adjudication processes (n=11, 26.2%). Sensitivity analysis excluding studies at high risk of bias showed similar pooled estimates, suggesting robust findings.

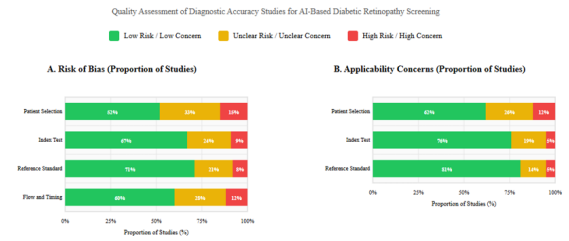

Figure 4: QUADAS-2 Risk of Bias and Applicability Concerns Assessment

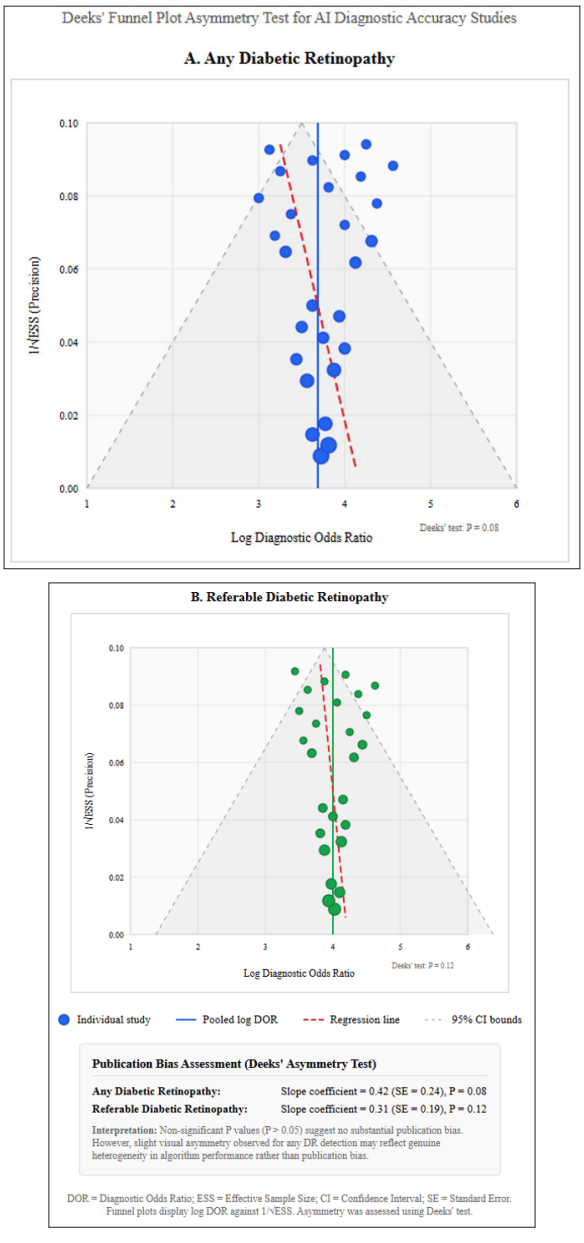

Figure 5: Funnel plot for Publication Bias
a. Any diabetic retinopathy
b. Referable Diabetic Retinopathy

Publication Bias

Visual inspection of funnel plots revealed slight asymmetry (Figure 5), with a tendency toward studies reporting higher diagnostic accuracy. However, Deeks' test did not reach statistical significance for either any diabetic retinopathy (P=0.08), suggesting that substantial publication bias was unlikely. The observed asymmetry may reflect the genuine variation in algorithm performance across different validation settings and populations.

Discussion

Summary of Main Findings

This systematic review and meta-analysis provides comprehensive evidence that AI and deep learning systems demonstrate high diagnostic accuracy for diabetic retinopathy screening. The pooled sensitivity of 87.5% and specificity of 84.2% for detecting any diabetic retinopathy, combined with an AUROC of 0.92, indicate that these technologies can achieve diagnostic performance approaching that of human expert graders. For referable diabetic retinopathy, the even higher sensitivity of 91.8% suggests particular utility in identifying patients requiring specialist referral.

To contextualize these findings clinically, using Fagan nomogram analysis, a positive AI screening result increases the probability of diabetic retinopathy from a baseline prevalence of approximately 35% to approximately 85–90%. Conversely, a negative result reduces the post-test probability to approximately 5–8%. These predictive values demonstrate substantial clinical utility for both ruling in disease requiring referral and ruling out disease in low-risk populations.

Sources of Heterogeneity

Our analysis identified several key sources of heterogeneity in diagnostic performance. The most clinically important finding was the significant difference between internal and external validation studies, with external validation showing lower performance (AUROC 0.90 vs 0.94). This gap highlights potential overfitting to training data distributions and underscores the critical importance of rigorous external validation before clinical deployment.

Image quality emerged as an important determinant of diagnostic accuracy. Studies using higher-resolution images (>1000×1000 pixels) demonstrated superior performance, reflecting the importance of detailed visualization of retinal microvasculature and subtle lesions for accurate detection of early diabetic changes. Algorithm architecture also influenced performance, with ensemble methods and CNN-based approaches showing the highest diagnostic accuracy.

Clinical Implications

The clinical implications of these findings span multiple domains. First, AI-based systems can dramatically improve the efficiency and coverage of diabetic retinopathy screening programs by processing large volumes of fundus images rapidly and consistently. This capacity is especially valuable in rural or underserved areas where ophthalmologists are scarce, enabling primary care facilities to extend screening access to populations lacking specialist care.

Second, AI-enabled telemedicine represents a promising avenue for expanding screening access. Patients in remote areas can have fundus photographs captured locally using portable cameras, with images transmitted electronically for automated AI analysis. This approach eliminates geographic barriers to screening participation while ensuring that limited specialist resources are directed toward patients with the greatest clinical need.

Third, accurate AI-based severity grading can support evidence-based treatment decisions and optimize resource allocation. By providing consistent, objective disease staging, these systems can help prevent both overtreatment of mild disease and undertreatment of vision-threatening conditions. Studies have demonstrated that AI-based screening is more cost-effective than manual grading, potentially providing efficient medical services with reduced per-patient screening costs.

Limitations and Challenges

Despite considerable promise, several important limitations must be acknowledged. Current AI systems maintain measurable error rates, with false negative rates of approximately 12% and false positive rates of approximately 9%. False negatives may lead to delayed treatment of sight-threatening disease, while false positives generate unnecessary referrals and patient anxiety. These limitations highlight the ongoing need for human oversight in clinical deployment.

The "black box" nature of deep learning models presents challenges for clinical adoption. Current systems cannot provide reasoning behind their diagnostic conclusions, potentially limiting clinician trust and effective human-AI collaboration. Development of explainable AI techniques represents an important research priority for next-generation systems.

Legal and ethical considerations around liability, algorithmic bias, data privacy, and equitable access require careful attention. Current legal frameworks do not clearly address responsibility when AI systems contribute to diagnostic errors. Professional societies and regulatory agencies must develop clear guidelines for ethical AI deployment in medical diagnosis.

Study Limitations

This systematic review has several limitations. First, substantial heterogeneity across studies limited the precision of pooled estimates. Second, some studies lacked detailed classification of diabetic retinopathy subtypes, affecting evaluation across the full disease spectrum. Third, meta-regression analyses may not have fully captured patient-level variables such as diabetes duration and comorbidities. Fourth, the reference standard of expert grading has its own imperfections, potentially affecting measured AI performance. Finally, most AI models were self-developed with limited transparency regarding pre-training and learning parameters.

Future Directions

Several research priorities emerge from this review. Enhanced data collection including detailed patient demographics and clinical characteristics will enable more comprehensive performance evaluation. Large-scale multi-center validation studies will improve algorithm generalizability across diverse populations. Development of human-AI collaboration models that combine

ophthalmologist expertise with AI assistance can optimize work efficiency and medical resource utilization. Finally, establishing standardized imaging protocols and quality benchmarks will help maximize AI performance across clinical settings.

Conclusion

Summary of Key Findings

This systematic review and meta-analysis, encompassing 42 studies and over 520,000 retinal images, provides comprehensive evidence that artificial intelligence and deep learning systems achieve high diagnostic accuracy for diabetic retinopathy screening. The pooled sensitivity of 87.5% (95% CI: 85.0–90.0%) and specificity of 84.2% (95% CI: 80.5–87.8%) for detecting any diabetic retinopathy, combined with an area under the receiver operating characteristic curve (AUROC) of 0.92, demonstrate that these technologies can achieve diagnostic performance approaching that of human expert graders.

For referable diabetic retinopathy—the clinically critical threshold for specialist referral—AI systems demonstrated even stronger performance, with pooled sensitivity of 91.8% (95% CI: 89.2–94.3%) and specificity of 87.5% (95% CI: 84.3–90.7%), yielding an AUROC of 0.95. These findings indicate that AI-based screening tools are particularly effective at identifying patients who require urgent ophthalmological evaluation, which is the primary goal of population-based screening programs.

Convolutional neural network architectures, particularly ensemble methods combining multiple models, demonstrated the highest diagnostic accuracy among the AI approaches evaluated. The widespread adoption of transfer learning techniques, where models pre-trained on large general image datasets are fine-tuned for diabetic retinopathy detection, has contributed to achieving high performance even with relatively limited ophthalmological training data.

Clinical Implications and Recommendations

The clinical implications of these findings are substantial and multifaceted. First, AI-based screening systems show considerable promise for addressing the growing global burden of diabetic retinopathy, particularly in regions facing acute shortages of trained ophthalmologists and retinal specialists. The World Health Organization estimates that approximately 35% of the 537 million adults with diabetes worldwide have some degree of

diabetic retinopathy, yet the majority lack access to regular screening. AI-enabled screening could dramatically expand coverage by enabling point-of-care screening at primary healthcare facilities, community health centers, and even pharmacies equipped with portable fundus cameras.

Second, our findings support the integration of AI systems as a triage tool within existing screening pathways. Rather than replacing human expertise, AI can serve as an efficient first-line filter, rapidly identifying patients with normal or mild disease who do not require immediate specialist attention, while flagging those with referable disease for priority review. This hybrid approach optimizes the allocation of scarce specialist resources while maintaining high-quality care standards.

Third, the telemedicine applications of AI screening are particularly compelling. Patients in remote or underserved areas can have fundus photographs captured locally using portable cameras, with images transmitted electronically for automated AI analysis and, when necessary, remote specialist review. This approach eliminates geographical barriers to screening participation and enables more frequent monitoring for high-risk patients without overwhelming specialist capacity.

However, several important caveats must guide clinical implementation. Our meta-analysis revealed that external validation studies demonstrated significantly lower diagnostic accuracy compared to internal validation (AUROC 0.90 vs. 0.94, $P<0.001$). This performance gap highlights the critical importance of validating AI systems on populations representative of intended deployment settings before clinical implementation. Algorithms trained predominantly on images from specific ethnic groups, camera systems, or clinical settings may show reduced performance when applied to different populations or imaging conditions.

Recommendations for Implementation

Based on our findings, we propose the following recommendations for healthcare systems considering AI implementation for diabetic retinopathy screening:

For clinicians and healthcare providers: AI systems should be implemented as decision-support tools rather than autonomous diagnostic systems. Human oversight remains essential, particularly for borderline cases, poor-quality images, and patients with complex presentations. Clinicians should understand the strengths and limitations of AI systems,

including the types of errors they tend to make, to appropriately interpret and act on AI-generated recommendations.

For policymakers and health system administrators: Implementation of AI screening should be preceded by rigorous local validation studies to confirm performance in the target population. Regulatory frameworks should be established to ensure appropriate oversight, quality assurance, and clear accountability for diagnostic decisions. Investment in imaging infrastructure, connectivity, and workforce training is essential to realize the full potential of AI-enabled screening programs.

For researchers and technology developers: Future development efforts should prioritize improving generalizability across diverse populations, imaging systems, and clinical settings. Development of explainable AI techniques that can provide clinicians with interpretable reasoning for diagnostic classifications will enhance trust and facilitate effective human-AI collaboration. Prospective studies evaluating real-world clinical outcomes, cost-effectiveness, and patient acceptability are needed to build the evidence base for widespread adoption.

Implications for Global Health Equity

The potential of AI to address global disparities in diabetic retinopathy screening access deserves particular emphasis. Low- and middle-income countries bear a disproportionate burden of diabetes-related blindness due to limited specialist availability, inadequate screening infrastructure, and financial barriers to care. AI-based screening offers a pathway to democratize access

to early detection, enabling timely intervention that can prevent irreversible vision loss.

However, realizing this potential requires deliberate attention to equity considerations. AI systems trained primarily on images from high-income country populations may perform less well when applied to underrepresented ethnic groups or in settings with different camera systems and imaging protocols. Ensuring equitable benefit from AI technology will require inclusive dataset development, local validation, and adaptation of implementation models to diverse healthcare contexts.

Limitations and Areas for Caution

Despite the encouraging findings of this meta-analysis, several limitations warrant caution in interpreting and applying these results. Current AI systems maintain measurable error rates, with false negative rates of approximately 12% and false positive rates of approximately 9% in pooled analyses. While these rates are comparable to human graders, they translate to clinically significant numbers of missed diagnoses and unnecessary referrals when applied at population scale.

The substantial heterogeneity observed across studies (I^2 exceeding 89% for both sensitivity and specificity) indicates that AI performance varies considerably depending on the specific algorithm, population, imaging conditions, and study methodology. This variability means that pooled estimates should not be uncritically assumed to apply in any given clinical context; local validation remains essential.

Legal and ethical frameworks for AI-assisted diagnosis remain underdeveloped in most jurisdictions. Questions of liability when AI systems contribute to diagnostic errors, appropriate standards for informed consent when AI is involved in clinical decisions, and mechanisms to ensure algorithmic fairness and prevent discrimination require ongoing attention from professional societies, regulators, and policymakers.

Future Research Priorities

Several key research priorities emerge from this systematic review. First, large-scale prospective studies evaluating AI screening in real-world clinical settings are needed to assess whether the high diagnostic accuracy observed in validation studies translates to improved patient outcomes, including reduced rates of vision loss and blindness. Second, head-to-head comparisons of different AI systems, ideally using standardized validation datasets, would help clinicians and health systems select optimal tools for their settings.

Third, research on optimal human-AI collaboration models is needed to determine how AI recommendations should be presented to clinicians, what confidence thresholds should trigger different clinical actions, and how to maintain appropriate human oversight without negating efficiency gains. Fourth, development and validation of AI systems capable of detecting multiple retinal pathologies simultaneously would enhance the value proposition of AI screening by identifying not only diabetic retinopathy but also glaucoma, age-related macular degeneration, and other sight-threatening conditions during a single screening encounter.

Fifth, economic evaluations comparing the cost-effectiveness of various AI implementation strategies across different healthcare contexts would inform resource allocation decisions. Such analyses should consider not only direct costs of AI systems and imaging equipment but also downstream effects on referral patterns, treatment costs, and productivity losses from vision impairment [1-25].

Concluding Remarks

In conclusion, this comprehensive systematic review and meta-analysis provides robust evidence that AI and deep learning systems have achieved sufficient diagnostic accuracy to serve as effective tools for diabetic retinopathy screening. With pooled AUROC values exceeding 0.90 for both any diabetic retinopathy and referable diabetic retinopathy detection, these technologies represent a significant advance in our capacity to identify and treat sight-threatening disease before irreversible vision loss occurs.

The demonstrated performance gap between internal and external validation underscores that successful clinical deployment requires rigorous validation in representative populations and careful attention to implementation factors that may affect real-world performance. AI should be viewed not as a replacement for human expertise but as a powerful augmentation that extends the reach and efficiency of specialist care.

The optimal approach combines AI technology with human clinical expertise, leveraging the complementary strengths of automated screening—consistency, scalability, and tireless processing capacity—with specialist diagnostic judgment, clinical context integration, and patient-centered care. This collaboration offers the most promising pathway to achieving the global goal of eliminating preventable blindness from diabetic retinopathy.

As the global diabetes epidemic continues to expand, particularly in low- and middle-income countries where specialist resources are most constrained, AI-based screening offers a transformative opportunity to bridge the gap between screening need and specialist capacity. Realizing this potential will require sustained investment in research, infrastructure, workforce training, and health system adaptation. The evidence synthesized in this review provides a strong foundation for evidence-based decision-making as healthcare systems worldwide navigate the integration of AI into diabetic retinopathy screening programs.

Ultimately, the success of AI in diabetic retinopathy screening should be measured not by algorithmic performance metrics alone but by its contribution to reducing the burden of preventable vision loss and blindness among people with diabetes worldwide. By enabling earlier detection, more efficient resource allocation, and expanded access to screening, AI technology has the potential to fundamentally improve outcomes for the hundreds of millions of people at risk of diabetic eye disease. The findings of this systematic review support continued investment in this promising technology while emphasizing the need for thoughtful, evidence-based implementation approaches that prioritize patient safety and health equity.

Acknowledgments
We acknowledge the contributions of all research team members involved in data extraction and quality assessment. No funding was received for this systematic review.

Data Availability Statement
The data supporting the findings of this systematic review are available from the corresponding author upon reasonable request. The complete search strategy and list of included studies are provided in the supplementary materials.

Author Contributions
NS and TT conceived and designed the study. NS, TT, and SL performed the literature search and study selection. NS and NV extracted data and assessed study quality. UV performed statistical analyses. NS wrote the first draft of the manuscript. All authors critically revised the manuscript for important intellectual content and approved the final version.

Conflict of Interest Statement
The authors declare no conflicts of interest related to this systematic review and meta-analysis.

Funding
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

1. Yau JW, Rogers SL, Kawasaki R, Lamoureux EL, Kowalski JW, et al. Global prevalence and major risk factors of diabetic retinopathy. Diabetes Care. 2012. 35: 556-564.
2. Wong TY, Sun J, Kawasaki R, Ruamviboonsuk P, Gupta N, et al. Guidelines on diabetic eye care: the International Council of Ophthalmology recommendations for screening, follow-up, referral, and treatment based on resource settings. Ophthalmology. 2018. 125: 1608-1622.
3. Ting DSW, Cheung CY, Lim G, Tan GSW, Quang ND, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA. 2017. 318: 2211-2223.
4. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016. 316: 2402-2410.
5. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med. 2018. 1: 39.
6. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015. 521: 436-444.
7. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011. 155: 529-536.
8. McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. JAMA. 2018. 319: 388-396.
9. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, et al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol. 2005. 58: 982-990.
10. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol. 2005. 58: 882-893.
11. Early Treatment Diabetic Retinopathy Study Research Group. Grading diabetic retinopathy from stereoscopic color fundus photographs--an extension of the modified Airlie House classification. ETDRS report number 10. Ophthalmology. 1991. 98: 786-806.
12. Wilkinson CP, Ferris FL 3rd, Klein RE, Lee PP, Agardh CD, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology. 2003. 110: 1677-1682.
13. Krause J, Gulshan V, Rahimy E, Karth P, Widner K, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology. 2018. 125: 1264-1272.
14. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. 770-778.
15. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z, et al. Rethinking the inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. 2818-2826.
16. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. 2019. 1: e271-e297.
17. Tufail A, Rudisill C, Egan C, Kapetanakis VV, Salas-Vega S, et al. Automated diabetic retinopathy image assessment software: diagnostic accuracy and cost-effectiveness compared with human graders. Ophthalmology. 2017. 124: 343-351.
18. Ting DSW, Pasquale LR, Peng L, Campbell JP, Lee AY, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol. 2019. 103: 167-175.
19. Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. Ophthalmology. 2017. 124: 962-969.
20. Bellemo V, Lim ZW, Lim G, Nguyen QD, Xie Y, et al. Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. Lancet Digit Health. 2019. 1: e35-e44.
21. Wang Z, Li Z, Li K, Mu S, Zhou X, et al. Performance of artificial intelligence in diabetic retinopathy screening: a systematic review and meta-analysis of prospective studies. Front Endocrinol. 2023. 14: 1197783.
22. Yip MYT, Lim G, Lim ZW, Nguyen QD, Chong CCY, et al. Technical and imaging factors influencing performance of deep learning systems for diabetic retinopathy. NPJ Digit Med. 2020: 3: 40.
23. Grauslund J. Diabetic retinopathy screening in the emerging era of artificial intelligence. Diabetologia. 2022. 65: 1415-1423.

24. Gunasekeran DV, Ting DSW, Tan GSW, Wong TY. Artificial intelligence for diabetic retinopathy screening, prediction and management. Curr Opin Ophthalmol. 2020. 31: 357-365.
25. Teo ZL, Tham YC, Yu M, Cheng CY, Wong TY, et al. Do we have enough ophthalmologists to manage vision-threatening diabetic retinopathy? A global perspective. Eye. 2020. 34: 1255-1261.