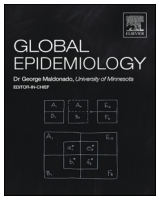
Global Epidemiology 9 (2025) 100209

Contents lists available at ScienceDirect

Global Epidemiology

journal homepage: www.sciencedirect.com/journal/global-epidemiology



Original Research

Artificial intelligence for advancing eye care in resource-poor settings: Assessing the predictive accuracy of an AI-model for diabetic retinopathy screening in India

Rohan Chawla [a], Prachi Karkhanis [b], Malay Shah [b], Aritra Das [b], Rishabh Sharma [b], Dhwani Almaula [b], Pradeep Venkatesh [a,*], Harsh Vardhan Singh [b], Mukul Kumar [b], Ramanuj Samanta [d], Vinod Kumarl [a], Amar Shah [c], Bhavin Vadera [c], Nakul Jain [b], Akanksha Sen [b], Shyamsundar Shreedhar [b], Vipin Garg [b], Soma Dhaval [b], Kowshik Ganesh [b], Srinivas Rana [b], Radhika Tandon [a]

[a] *All India Institute of Medical Sciences, Ansari Nagar, New Delhi 110029, India*
[b] *Wadhwani AI, LORDS EDUCATION & HEALTH SOCIETY (LEHS), 70, Ring Road, 2nd Floor, Lajpat Nagar-III, New Delhi 110024, India*
[c] *United States Agency for International Development (USAID), India*
[d] *All India Institute of Medical Sciences, Rishikesh, Uttarakhand 249203, India*

A B S T R A C T

Background: Timely identification and treatment of Diabetic Retinopathy (DR) is critical in avoiding vision loss. DR screening is challenging, especially in resource-limited areas where trained ophthalmologists are scarce. AI solutions show promise in addressing this challenge. In this study, the performance metrics of an AI solution (MadhuNetrAI) developed in India was evaluated for referring and grading DR.

Methods: MadhuNetrAI was developed de novo by the All India Institute of Medical Sciences (AIIMS) and Wadhwani AI (WIAI). It was tested on 1078 fundus images (from AIIMS Delhi and an unannotated subset of publicly available EyePACS images) against two ophthalmologists and an adjudicator serving as independent gold-standard annotators, wherein the disease status of the patients remained unknown.

Findings: MadhuNetrAI demonstrated high sensitivity (93·2 %; CI: 89·5 %–95·6 %) and specificity (95·3 %; CI: 93·7 %–96·6 %) in detecting referable DR (moderate, severe, proliferative DR). The area-under-the-curve for referring DR against the gold standard was 0·97 (CI: 0·95–0·99) indicating excellent diagnostic performance. The agreement in grading DR severity was high (kappa = 0·89, CI: 0·86–0·91). The model performed comparably in detecting DR too.

Interpretation: MadhuNetrAI's ability to grade DR severity and identify referrable cases could bring DR patients to care much earlier. Further research and clinical trials are needed to ensure its reliability and generalizability across diverse populations and image qualities.

Funding: MadhuNetrAI was developed by technical and programmatic teams at WIAI, with inputs and contributions by the clinical team at AIIMS, and funded by USAID. The authors have no financial or non-financial conflicts of interest to disclose.

Introduction

Diabetic retinopathy (DR) is a severe microvascular complication of diabetes mellitus (DM), particularly for individuals with prolonged or poorly managed diabetes, and one of the leading causes of vision loss worldwide associated with blindness in almost four million people [1]. The prevalence of DR in South-East Asia Region (SEAR), has been reported to vary across different regions, depending on factors such as the degree of urbanization, the presence of migrant populations, and racial demographics [2,3]. In India, where over 77 million people are afflicted by diabetes, the prevalence of DR has reached alarming levels, with approximately three million people aged 40 years and above having vision-threatening DR [4]. The universally accepted strategy to curb this public health problem is to implement an efficient screening strategy with easily accessible diagnostic methods preventing visual impairment in patients with diabetes [5].

Traditionally, ophthalmologists rely on fundus examinations or fundus photography to diagnose DR by evaluating retinal images [6]. However, these methods are heavily dependent on the availability of trained ophthalmologists, especially for grading DR severity.

* Corresponding author at: Centre for Ophthalmic Sciences, AIIMS Campus Temple, Sri Aurobindo Marg, Ansari Nagar, Ansari Nagar East, New Delhi, Delhi 110029, India.
E-mail address: venkyprao@yahoo.com (P. Venkatesh).

https://doi.org/10.1016/j.gloepi.2025.100209
Received 28 November 2024; Received in revised form 31 May 2025; Accepted 2 June 2025
Available online 4 June 2025

Consequently, a significant portion of the population, particularly in low- and middle-income countries (LMICs), face challenges in accessing timely DR detection and management. This is especially relevant in SEAR, where only 4 out of 10 reported countries have achieved an adequate ophthalmologist-population ratio (1:100,000) [7]. Moreover, the urban-rural distribution is highly disproportionate, with most ophthalmologists located in urban areas [7]. In a country like India, where a vast majority of the population resides in rural areas, the lack of trained ophthalmologists creates an undue burden on the health institutions (such as tertiary care centres and speciality hospitals) where such facilities are available, making timely DR detection and grading a formidable challenge. This problem can be overcome by instituting a screening and referral mechanism in rural and poorly connected communities that does not depend on the need for a trained ophthalmologist.

Addressing this screening gap is imperative, and artificial intelligence (AI) holds great promise in addressing the diagnostic obstacles posed by DR in resource-limited settings [8]. AI algorithms have demonstrated remarkable abilities to analyze medical images and provide accurate diagnoses [9,10]. Neural network-based diagnostic algorithms have been successfully tested for detecting retinal pathologies as well, e.g. retinal layer segmentation [11], intraretinal fluid in case of macular oedema [12,13], and, in general, for analyzing Optical Coherence Tomography (OCT) scan images [14]. AI, especially deep learning, has enabled the development of autonomous screening tools for DR as well achieving accuracy levels comparable to human experts in some cases [15,16]. For instance, Idx-DR, an AI model approved by the FDA, demonstrated high sensitivity and specificity in detecting referable DR and is widely recognized for its real-world clinical application, setting a precedent for AI's potential to scale DR screening efforts [17]. Past studies have also shown that AI-driven models such as those developed by EyePACS and Google have achieved over 90 % sensitivity and specificity in detecting referable DR across diverse datasets, including the Messidor-2 and EyePACS-1 datasets [15,16]. Thus, utilizing AI for detecting and grading DR can enhance access to timely and dependable assessments, bridging the gap between patient needs and the limited availability of ophthalmologists [18]. Such AI solutions can serve as a valuable tool for timely detection, particularly in resource-poor settings like India, where widespread adoption of smartphones and portable imaging devices make telemedicine a viable option [19].

Despite promising results, AI-driven DR screening tools face several challenges. One primary limitation is variability in performance due to differences in image quality and device types used for fundus photography. For instance, while models like Idx-DR and EyeArt have demonstrated effectiveness in controlled environments, real-world deployment often presents challenges in maintaining consistent image quality across different camera types [17,18]. Furthermore, ethical concerns regarding fairness and bias in AI systems, especially when deployed in low- and middle-income countries (LMICs), necessitate more robust training datasets that capture diverse demographic and clinical characteristics [20,21].

Against this backdrop, an indigenous AI solution (MadhuNetrAI) was developed by Wadhwani AI (WIAI), the AI Unit of the Ministry of Health and Family Welfare (MoHFW), Government of India (GoI). The AI model is based on a Convolutional Neural Network (CNN) backbone that ingests the retinal fundus image to produce an encoded representation of the fundus and its features, which are then ordinally regressed to obtain a DR grade as its output. To our knowledge, this is the first AI solution developed in India that determines not only referable cases of DR, but also the severity grade of DR according to the International Classification of Diabetic Retinopathy as *No Apparent DR* (Grade 0), *Mild Non-proliferative DR* (Grade 1), *Moderate Non-proliferative DR* (Grade 2), *Severe Non-proliferative DR* (Grade 3), and *Proliferative DR* (Grade 4) [22]. This development addresses a critical gap in the Indian healthcare landscape, promising to enhance timely detection and intervention strategies for DR.

Development of the AI solution

The development stages of the AI solution (*MadhuNetrAI*) for screening DR encompassed several crucial phases. Initial data sourced 42,967 fundus images from open-source datasets EyePACS [23] (n = 35,126), APTOS [24] (n = 3662), and ODIR-5k [25] (n = 4179). The data was evenly split across these datasets in a ~80:10:10 ratio for model training, validation, and testing. Model training involved iterative optimization of the model's learnable parameters to satisfy the objective of grading DR severity. The validation set aids in tuning the hyperparameters of the model to achieve optimal performance. Finally, the test set is used to evaluate the efficacy of the trained model. The process flow envisioned for the AI solution for the screening of DR is depicted in Fig. 1.

As with every screening tool, the clinical utility of this AI solution needs to be rigorously evaluated and compared to established diagnostic standards to ensure its reliability and safety across different settings. This study presents an inaugural evaluation of the AI solution for screening DR by the All India Institute of Medical Sciences (AIIMS), with input from experienced ophthalmologists to assess the tool's performance in this context.

Further, evaluating AI solutions on images from different kinds of cameras is essential to ensure that these technologies are device-agnostic, demonstrating their effectiveness and reliability across a broad spectrum of imaging equipment. Several studies from India have validated device-agnostic AI for DR screening, demonstrating effectiveness across traditional fundus cameras and smartphone-based devices [8,26]. Although these reports highlight AI's potential to enhance accessible and scalable eye screening solutions, they are often developed and disseminated through commercial channels. In contrast, MadhuNetrAI was developed as a not-for-profit initiative and is not commercially marketed. The present manuscript signifies the first installment in a series dedicated to evaluating the performance of the AI solution from diverse sources.

Study objective

The primary objective of the study was to assess the predictive accuracy of an AI solution (MadhuNetrAI) to classify fundus images for DR, as referable and non-referable. The study aimed to benchmark the AI solution's predictive performance against the established gold standard (GS) annotation provided by expert ophthalmologists. Secondary objectives were to study the performance metrics of the model in grading the severity of DR and to assess the inter-rater variability between the annotations performed by two ophthalmologists and that between the humans and AI solution.

Materials and methods

Ethics approval

The use of data for this study received approval from the Institutional Review Board at AIIMS Delhi (approval reference number is IEC-628/05·08·2022, OP-24/06·10·2023). Ethical considerations encompassed patient confidentiality and adherence to established ethical principles and standards for human research in accordance with the Declaration of Helsinki.

Study design

The study adopted a blinded comparison design to evaluate the AI solution's predictive accuracy. A diverse set of fundus images was evaluated. Prospectively, each image was independently assessed by the AI solution and two expert ophthalmologists ('annotators') from AIIMS Delhi with at least five years of domain expertise. The annotators assigned one of the five internationally recognized grades of DR to each

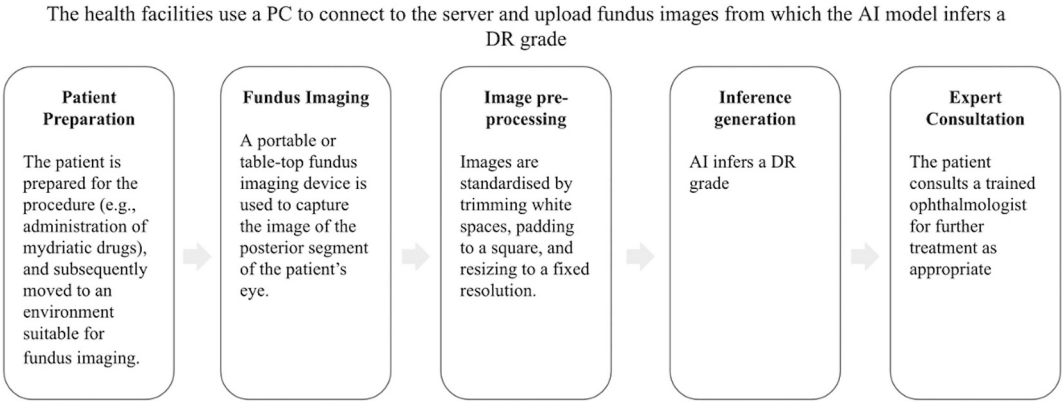


Fig. 1. Proposed process flow of screening patients for Diabetic Retinopathy using the AI solution.

fundus image. Any grade above Zero was considered DR (mild, moderate, severe, proliferative DR), and grades above One were considered 'referrable, which includes moderate, severe, proliferative DR'. The diagnosis provided by the annotators served as the reference GS annotation. If there was any disagreement between the two annotators regarding any fundus image – either related to the diagnosis/grading or its inclusion in the study – a senior ophthalmologist (~20 years of domain expertise), the 'adjudicator', evaluated the same image, and the opinion was considered final. The annotators and adjudicator underwent rigorous training on the annotation protocol to ensure standardization. They independently conducted the annotation of the fundus images using an exclusive password-protected online tool specifically developed for the purpose. They were also blinded to the demographic and clinical information of the patients (e.g., the age of the patient, whether the patient was diabetic or not etc.). The Standards for Reporting of Diagnostic Accuracy (STARD) guidelines were followed [27].

Data source

The study involved a retrospective analysis of dilated posterior pole fundus images sourced from the Department of Ophthalmology (Rajendra Prasad Centre for Ophthalmic Sciences) at AIIMS Delhi and publicly available subset of unannotated images from EyePACS. All images captured at AIIMS were macula-centered and obtained using a non-mydriatic Zeiss tabletop fundus camera (Model: FF450 Plus IR) with a 50° field of view, ensuring uniformity in image quality and characteristics for this subset of data. However, for the images sourced from EyePACS, the specifications of the capturing devices including the field of view and type of camera were not available. As EyePACS is an open-source dataset comprising images collected in diverse clinical settings, the possibility of variations in the type of retinal capturing devices may not be overruled. These images collectively, without patient identifiers, provided the foundation for assessing the AI solution's performance. In total, 3117 fundus images were reviewed (2339 from AIIMS and 778 from EyePACS), of which 1078 (34·6 %; 555 from AIIMS and 523 from EyePACS) were used in the final analysis. The process of reviewing fundus images to determine inclusion in the evaluation study is depicted via a flowchart in Fig. 2. Only images that could be reliably graded by humans, including those with slight haziness, minimal vitreous haemorrhage, or cataracts, were included in the study. The criteria for excluding fundus images were: lack of consent from the patient regarding the use of their fundus images, presence of significant media opacities/artefacts, other ophthalmologic pathologies and poorly focused images that affected the ability to identify the retinal lesions. Based on the above criteria, a total of 704 images were excluded from the evaluation study.

Sample size

Given the uncertainty over the expected prevalence of DR among the obtained fundus images, we could not perform a sample size estimation for the diagnostic test accuracy. Nevertheless, we conducted a post-hoc power calculation based on the proportion of fundus images that were identified as having DR by the GS annotators (30 %). We also assumed a sensitivity and specificity of 80 % each, an alpha (α) error of 5 %, and the Farrington-Manning Score Test, which assumes asymptotic normality for comparison [28]. Based on the aforementioned parameters, the post-

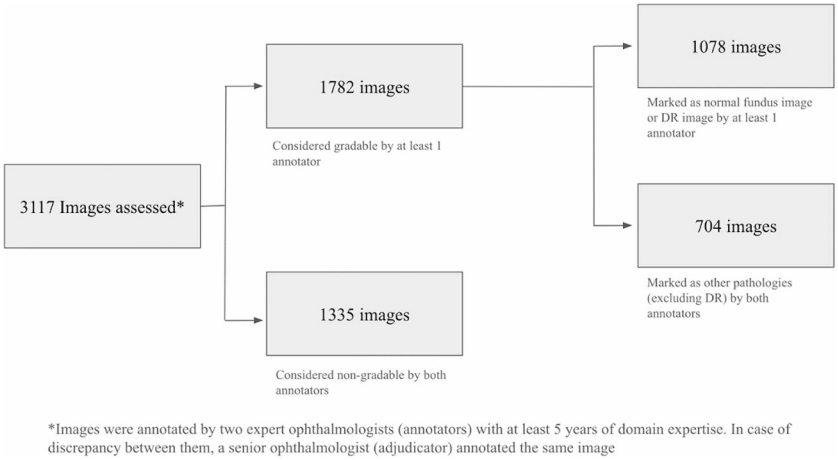


Fig. 2. Flowchart depicting the selection of fundus images for evaluation.

facto statistical power (1 – β error) for sensitivity and specificity on a sample of 1078 fundus images turned out to be 72 % and 100 %, respectively. The power calculation was performed using the *Proc Power* procedure in SAS v9·4.

Statistical analysis

Statistical analyses included calculating sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) for the AI-predicted DR grades against the GS annotations for detecting and for determining referable cases of DR. Cohen's simple kappa statistic was employed to evaluate the inter-rater agreement on DR diagnosis (present/absent) between the AI solution and the GS annotation. We also analyzed the receiver operating characteristic (ROC) curve to determine the area under the curve (AUC). The same analyses were repeated for 'referable' DR (grades Zero or One vs grades Two, Three or Four).

The overall predictive accuracy regarding the exact grade of DR was assessed by percent agreement (calculated as the ratio of the number of times the AI solution agrees with the GS annotation to the total number of annotations), and Cohen's quadratic weighted kappa statistic (used as a conservative estimate since it factors in the agreement that occurred by chance). The level of agreement as per simple or quadratic weighted kappa statistic was interpreted as per the following scale: slight agreement = up to 0·2, fair ≥0·2–0·4, moderate ≥0·4–0·6, substantial ≥0·6–0·80, and almost perfect agreement ≥0·8 [29,30].

We also compared the agreement (or disagreement) between the diagnoses provided by the GS. This approach served to gauge the reproducibility of screening results across evaluations by expert ophthalmologists, thus indirectly assessing the precision of the AI tool. This proxy comparison allowed us to approximate the reliability of the diagnostic outputs without conducting a formal test/retest procedure.

We also assessed the correlation in DR grading between the AI solution and the GS annotation using Spearman's Rank Correlation test. For these comparisons, a *p*-value less than or equal to 0·05 was considered significant. All statistical analyses were performed using SAS v9·4.

Role of the funding source

The funding agency of the study had no role in the study design, data collection, data analyses, data interpretation, or writing of the report. The authors have no financial or non-financial conflicts of interest to disclose.

Results

In total, 1078 fundus images were determined gradable by the annotators and used for evaluation. Regarding the establishment of the GS annotation, the two grading annotators had an agreement on 846 (78·5 %) of the fundus images (for both the diagnosis of DR and its grading). For the rest of the 232 images (21·5 %), consultation was sought with the adjudicator, and their opinion was considered final. The summary of annotation findings on DR diagnosis is presented in Fig. 3.

Of the evaluated fundus images, approximately one-third (n = 324, 30·1 %) were determined to be DR (Among the 555 images from AIIMS, 228 were DR; among the 523 images from EyePACS, 96 were DR) by the ophthalmologists, whereas 32·4 % (n = 349) were predicted as DR by the AI solution. The sensitivity, and specificity (and 95 % Confidence Interval (CI)) of the AI solution in determining the presence of DR against the gold standard annotation are 90·1 % (86·9 %–93·4 %), and 92·4 % (90·6 %–94·3 %) respectively (Table 1). Moreover, the simple kappa coefficient of 0·81 (95 % CI: 0·77–0·85) suggested an 'almost perfect' agreement between the AI solution and the gold standard annotators for detecting DR.

The Youden's index and overall predictive accuracy of the AI solution in determining the presence of DR were also reasonably high, at 0·83 and 91·7 % (95 % CI: 89·9 %–93·3 %), respectively. The AUC from ROC curve analysis (Fig. 4) was 0·95 (95 % CI: 0·93–0·97), indicating that the model is highly capable of determining the presence/absence of DR. A Somers' D value of 0·83 (95 % CI: 0.79–0.86) suggests that the AI solution predicted the gold standard annotations very well. The sensitivity of 93·2 % (95 % CI: 89.5 %–95.6 %), specificity of 95·3 % (95 % CI: 93.7 %–96.6 %), and overall predictive accuracy (94·8 % (95 % CI: 93.3 %–96.0 %) of the AI solution in determining the referable cases of DR are depicted in Supplementary Table 1.

A sensitivity analysis was performed by adjusting the referral threshold to evaluate the model performance where only severe non-proliferative DR and proliferative DR were considered as referable cases. The model achieved a sensitivity of 89.3 % (95 % CI: 77.1 %–90.7 %) and specificity of 95.9 % (95 % CI: 94.6 %–97.1 %) in this case. When only PDR cases were considered for referral, the sensitivity dropped to 55.9 % (95 % CI: 45.0 %–66.0 %) and specificity 99.0 % (95 % CI: 98.3 %–99.6 %). These findings suggest that while focusing solely on proliferative DR reduces the number of referrals substantially, it may risk missing a significant number of severe cases as depicted in Fig. 5. This limitation is likely attributable to the smaller sample size, and improvements in this metric are observed as the sample size increases (refer to Supplementary Document 1). To the best of our knowledge, there are no published discussions addressing this surprising observation. It is possible that, for this reason, even AI models currently provide only a basic classification — detecting DR or categorizing cases as referable or non-referable DR. We conjecture that a key reason for the model showing subpar performance in patients with proliferative DR, would be the presence of only fine abnormal blood vessels (which is indicative of proliferative DR) in the absence of other features of proliferative DR such as extensive and well-formed vascular networks, extensive retinal haemorrhages, preretinal haemorrhage, mild vitreous haemorrhage etc.

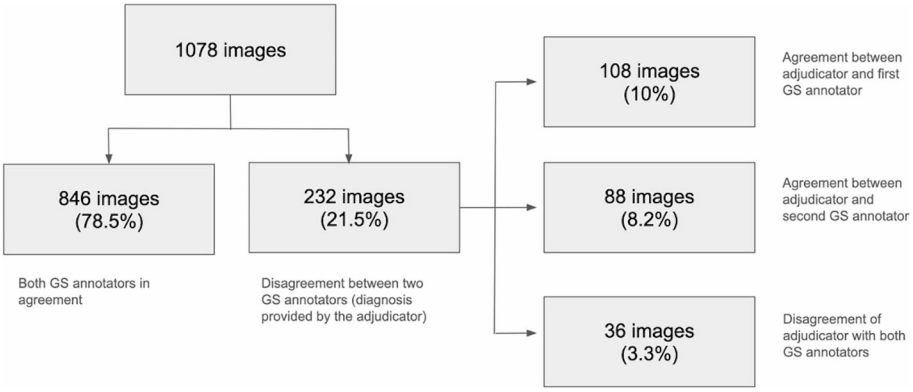


Fig. 3. Summary of annotation findings to establish the gold standard for diabetic retinopathy diagnosis (N = 1078).

Table 1
Performance of the AI model in determining the presence of diabetic retinopathy ($N = 1078$).

		Gold standard annotators' (Ophthalmologists') diagnosis of presence of DR						Total
		Yes			No			
		n	Row (%)[a]	Column (%)[b]	n	Row (%)[a]	Column (%)[b]	
AI prediction on presence of DR	Yes	292	83·7 %	90·1 %	57	16·3 %	7·6 %	349
	No	32	4·4 %	9·9 %	697	95·6 %	92·4 %	729
		324			754			1078

[a] % out of all AI predictions.
[b] % out of all gold standard annotation diagnoses.

The high value of the kappa coefficient (0·86; 95 % CI: 0·83–0·90) and AUC of 0·97 (95 % CI: 0·95–0·99) (Supplementary Fig. 1) also suggested almost perfect agreement between the GS annotators and AI solution. For both the models (DR detection and referrable DR), the operating point of the model can potentially be chosen in consultation with public health experts to balance sensitivity and specificity. This will ensure that the model can identify as many cases of DR as possible without producing too many false positives.

Regarding the grading of DR, the quadratic weighted kappa coefficient was 0·89 (95 % CI: 0·86–0·91), and the overall predictive accuracy compared to the gold standard annotation was 81·0 % (95 % CI: 78.6 %–83.3 %). The performance of the AI solution in determining the grade of DR is presented in Table 2. The Cochran Mantel Haenszel (CMH) statistic of the association between disease classification by the AI and the gold standard annotation was strongly significant (CMH statistic: 850; p-value <0·0001). This finding is also corroborated by a high coefficient in Spearman Rank correlation (rho (ρ): 0·87 (95 % CI: 0.85–0.88); p-value <0·0001).

Tables 3 and 4, respectively, depict the extent of agreement on DR diagnosis and DR grading between the gold standard annotators. Cohen's simple kappa coefficient for DR diagnosis was 0·85 (95 % CI: 0·82–0·89) and the quadratically weighted kappa coefficient for DR grading was 0·92 (95 % CI: 0·90–0·94).

Discussion

This study provides a structured framework for the evaluation of an AI solution's (MadhuNetrAI) accuracy in classifying DR images as referable/non-referable and unlike other reported models, also in grading the severity of DR and its level of agreement with human experts. Our observations shed light on its potential role in DR screening and patient care, as well as limitations in regard to classification of higher severity grades. The measures of predictive accuracy were notably high, signifying the AI solution's robustness in correctly detecting the presence or absence of DR. Further, we found that the inter-rater agreement on DR detection and grading between expert ophthalmologists (GS annotators) was of a similar extent to that between the AI solution and the GS annotations.

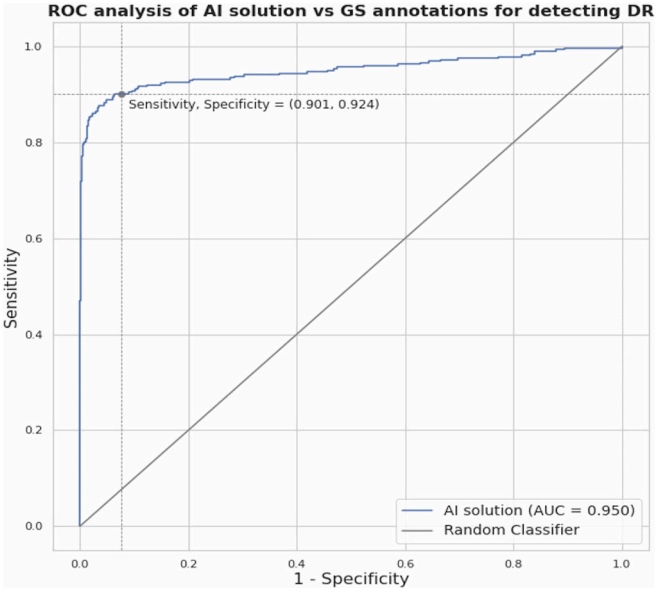


Fig. 4. ROC curve analysis on the predictive capacity of the AI model to detect diabetic retinopathy (N = 1078).

These findings are significant for several reasons. First, they suggest that the AI solution has the potential to overcome the challenges of DR screening in settings, where access to ophthalmologists is limited [7]. Second, while some limitations were observed in detecting the severity grade in most advanced stages, the AI solution's overall performance remains comparable to other AI-based [15–17] DR screening solutions in detecting both DR and referrable cases, which is important for ensuring that patients receive the appropriate treatment and care. Third, the high inter-rater agreement between the AI solution and the GS annotation suggests potential for assisting in DR screening workflows. Finally, it can be considered that the discrepancy between the AI solution and the GS annotation is not greater than the variation that exists among human experts. Given that the primary use of the AI solution will be for screening and subsequent referral of potential cases who would need advanced medical care, the performance of this AI solution in its current form can be considered adequate.

Importantly, we found that the accuracy measures performed even better when the referral threshold was set at Grade 2 and above, in accordance with the referral criteria for DR [31]. This suggests that aligning the AI solution with established international referral guidelines could enhance its real-world applicability and optimize patient triage for timely ophthalmological evaluation. Point-of-care DR screening could be particularly helpful in triaging two types of patients: those who do not need specialist referrals and those with referrable DR [32]. In the current study, as per the GS annotators, the proportion of patients who may have required referral was 24·5 %, not much different from a prior study that also evaluated an AI solution for DR [33]. The results assume a greater significance as our AI solution exhibits superior performance in determining the cases that need referrals. Successful screening, through automated analysis of fundus images, of such 'referrable' patients would allow for accelerated treatment and prevention of potentially vision-threatening disease. Another strength of the AI solution is that it is able to grade DR severity in addition to detecting its presence. This is important because the severity of DR determines the ability to differentiate referrable from non-referable DR along with the type of treatment and monitoring that patients require.

The implications of this AI tool extend far beyond the evaluation context, especially in India, where the burden of diabetes and its ocular complication, DR is substantial. AI-based screening models have demonstrated strong potential to address critical shortages in trained ophthalmologists by automating image analysis with high sensitivity and specificity, thus offering a scalable solution to improve early detection rates in at-risk populations [15,16]. For instance, systems like Idx-DR, have demonstrated high sensitivity in detecting referable DR, reinforcing the feasibility of deploying AI tools at the primary care level

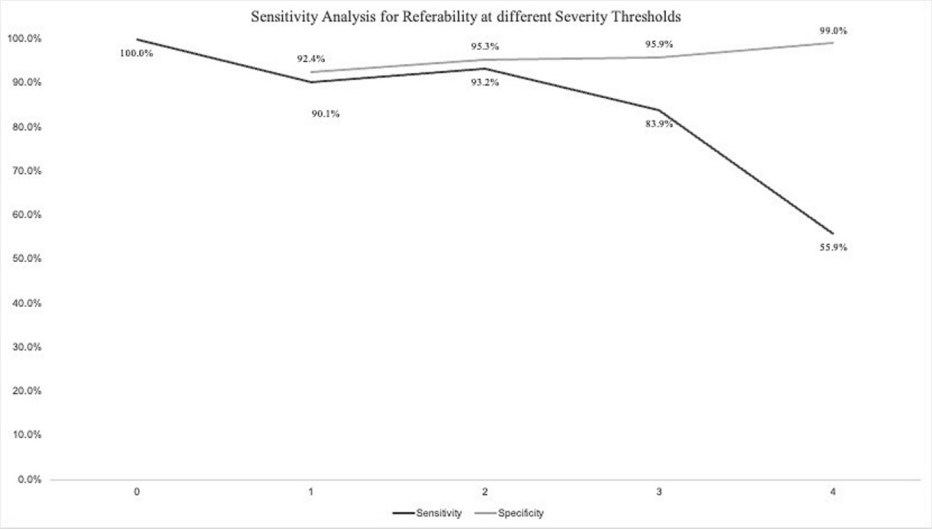


Fig. 5. Sensitivity analysis for referability of diabetic retinopathy at different thresholds (N = 1078).

Table 2
Performance of the AI model in determining the grade of diabetic retinopathy (N = 1078).

		Gold standard annotators' (Ophthalmologists') diagnosis of DR grade					
		DR absent (%)[a]	Mild NPDR (%)[a]	Moderate NPDR (%)[a]	Severe NPDR (%)[a]	Proliferative DR (%)[a]	Total
AI prediction of DR grade	DR absent (0)	697 (92·4)	23 (38·3)	5 (3·3)	0 (0)	4 (4·3)	729
	Mild NPDR (1)	45 (6·0)	11 (18·3)	6 (4·0)	0 (0)	3 (3·2)	65
	Moderate NPDR (2)	12 (1·6)	25 (41·7)	102 (67·1)	3 (15·8)	8 (8·6)	150
	Severe NPDR (3)	0 (0)	1 (1·7)	34 (22·4)	11 (57·9)	26 (28·0)	72
	Proliferative DR (4)	0 (0)	0 (0)	5 (3·3)	5 (26·3)	52 (55·9)	62
	Total	754	60	152	19	93	1078

NPDR – non-proliferative DR.
[a] % out of all gold standard annotations.

Table 3
Comparison of DR diagnoses made by two gold standard annotators (ophthalmologists) (N = 1078[a]).

		Annotator 1's diagnosis of presence of DR						Total
		Yes			No			
		n	Row (%)[b]	Column (%)[c]	n	Row (%)[b]	Column (%)[c]	
Annotator 2's diagnosis of presence of DR	Yes	267	89·3 %	90·2 %	32	10·7 %	4·8 %	299
	No	29	4·4 %	9·8 %	629	95·6 %	95·2 %	658
		296			661			

[a] 121 images were not graded (deemed ungradable or indeterminate) by either one of the annotators, but eventually deemed to be gradable by the adjudicator.
[b] % out of all diagnoses by annotator 2.
[c] % out of all diagnoses by annotator 1.

Table 4
Extent of (dis)agreement on DR grading between the gold standard annotators (N = 1078).

Grades of DR	Perfect agreement	Gold standard annotator	
		Annotator 1	Annotator 2
	[n (%)[a]]	[n (%)[a]]	[n (%)[a]]
DR Absent (0)	629 (58·3)	679 (63·0)	720 (66·8)
Mild NPDR (1)	27 (2·5)	54 (5·0)	63 (5·8)
Moderate NPDR (2)	115 (10·7)	160 (14·8)	146 (13·5)
Severe NPDR (3)	11 (1·0)	15 (1·4)	28 (2·6)
Proliferative DR (4)	64 (5·9)	89 (8·3)	75 (7·0)
Not graded[b]	–	81 (7·5)	46 (4·3)
Total graded	846 (78·5)	997 (92·5)	1032 (95·7)

[a] % out of all annotated fundus images (N = 1078).
[b] Deemed ungradable or indeterminate) by either annotator.

to reduce referral bottlenecks [17]. Similarly, EyeArt, validated in a large-scale study across multiple sites, achieved a sensitivity of 91 % and specificity of 88 %, reinforcing AI's capacity to identify referable DR cases effectively [20]. Furthermore, recent reviews underscore the importance of device-agnostic performance in AI solutions for telemedicine-based DR screening, allowing for broader adaptability across various imaging devices and settings [20,34]. In this evaluation, *MadhuNetrAI* demonstrated accuracy levels comparable to these established models, suggesting its viability as a robust, reliable tool for DR screening within the unique socio-economic landscape of India. Moreover, as noted in other studies, ensuring that AI models are trained on diverse datasets that represent the heterogeneity of LMIC populations is critical for enhancing generalizability and mitigating biases [21]. Our findings align with global trends in AI-driven healthcare interventions, which emphasize the need for adaptable, scalable solutions capable of bridging healthcare gaps, especially in regions facing a rapid rise in DR

incidence [20].

Another factor essential to evaluating the utility of AI models is their adaptability across different devices and clinical settings. While AI-based automated DR screening models have demonstrated high diagnostic performance, their implementation across a broad array of device types remains limited, particularly in resource-constrained environments [35]. However, in the primary care setting of the LMICs, it is crucial for an AI system to be device-agnostic, and user-friendly [8,26,34], as the intended setting and user population may consist of a variety of cameras, technicians and staff without any prior experience in retinal imaging. Given the simple steps required for implementing *MadhuNetrAI*, we believe it could perform well at primary or secondary care sites. As a way forward, we plan to evaluate the performance of the AI solution on different cameras used in LMICs including handheld, the implementation feasibility, and ease of use of the AI solution in primary and secondary care settings with staff having no prior retinal imaging experience.

Deploying AI in healthcare settings presents ethical and operational challenges, such as algorithmic bias and data representativeness, which are critical considerations for scalability and impact. While several AI models have shown effectiveness in DR screening, few have been validated on diverse patient populations, which can lead to potential biases [21]. In a healthcare landscape marked by high heterogeneity in terms of demographics and socioeconomic status, *MadhuNetrAI* stands out for being trained on a dataset that includes images from a wide array of sources, addressing some of the limitations identified in other models. By focusing on validating the model for varied Indian contexts, *MadhuNetrAI* aligns with current ethical frameworks that emphasize fairness and inclusivity in AI deployment [17]. This emphasis positions *MadhuNetrAI* as a tool not only for improved diagnostic accuracy but also as a potential agent of health equity.

However, despite its promise, there are inherent limitations. The AI solution's performance is contingent on the quality and diversity of the training data. Moreover, the performance may vary in diverse population groups and diverse image qualities, necessitating contextual fine-tuning and validation in varying clinical environments. For instance, the evaluation study obtained the fundus images from a tertiary care centre (AIIMS, New Delhi) and open-source data. This could limit the generalizability of the study findings as the predictive accuracy may vary in other settings, such as primary care clinics or remote areas. Additionally, a limitation of this dataset is that we do not know the disease status of the participants, as the dataset did not contain any associated metadata. Furthermore, we did not track whether both eyes from the same individual were included, which may affect the generalizability of the findings and the potential correlations between the eyes of the same patient. Further, we used single-field macula-centered imaging, which, while practical for large-scale deployment in resource-limited settings, may result in under-detection of peripheral DR lesions. However, such cases are likely few in the context of a screening program aimed at expanding access through low-cost devices and outreach to peripheral health centers. Another potential limitation, similar to other AI solutions, could be variations in the image quality and resolution depending on the imaging device used to capture the fundus images. Variations in image quality typically also happen due to varying degrees of patient cooperation and operator skills, as is often the case with assessments conducted with actual patients. Discarding the poor quality of images may have some bearing in real-world settings, where image quality can vary. Ungradable images, lower-quality images from mobile attachments, or differences in image centering (e.g., macula-centered vs. disc-centered) could lead to poor or unreliable model performance. We also acknowledge that the algorithm is currently trained only to detect DR. Further work is needed to focus on detecting other retinal disorders including diabetic macular edema. The model also shows reduced sensitivity at higher severity grades (3 and 4), increasing the risk of missing truly referable cases. This is likely due to stricter thresholding and the low prevalence of severe cases in the dataset, which limits the model's learning. These factors highlight the importance of careful threshold selection, with grade 2 potentially offering a better balance between sensitivity and specificity for effective screening. Lastly, a relatively modest sample size (1078 images) was used in the current evaluation. However, we have subsequently analyzed a larger dataset (2566 images) — a superset of the images analyzed in this study — which shows significant improvement in the model's performance metrics, sensitivity analysis, and performance across individual severity grades as presented in Supplementary Document 1, reinforcing the potential of the AI solution. Nevertheless, it is critical to recognize that effective DR screening can only be impactful when robust systems for referral and treatment are in place.

In conclusion, this evaluation study underscores the potential of the *MadhuNetrAI* solution in revolutionizing DR screening, especially in resource-constrained settings. The findings have great public health relevance and potentially important implications for the diabetic retinopathy screening program in India. While some reduction in sensitivity is noted at higher grades, this observation suggests opportunities for further refinement. Further validation through population-based studies diverse demographics and various types of cameras are imperative to ensure broader applicability and generalizability across the geographies.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gloepi.2025.100209.

Research in context

Evidence before this study

Diabetic retinopathy (DR) is one of the leading causes of vision impairment, especially in South-East Asia Region (SEAR) where countries like India, Bangladesh and Indonesia have the highest prevalence of diabetes. Early screening and referral for treatment can prevent irreversible blindness, but many countries in SEAR do not have access to trained ophthalmologists. Artificial Intelligence technologies have been leveraged to provide reliable analysis by analyzing medical images, including that for retinal pathologies. These solutions can improve access to timely assessments, especially in resource-limited settings like India, where telemedicine is becoming increasingly popular.

Added value of this study

The AI solution evaluated in this study showed high diagnostic accuracy and agreement with human experts for DR screening. These findings suggest that the AI solution can help overcome the challenges of DR screening in resource-poor settings and can detect the grades of DR accurately. Overall, the performance of this AI solution is adequate for screening and referring potential cases in need of advanced medical care.

Implications of all the available evidence

This AI solution has the potential in India and other resource-poor settings to help overcome the lack of access to specialized ophthalmic services by empowering healthcare providers in remote locations with tools to efficiently screen and refer potential cases for better care. By making DR screening more accessible and affordable, this AI solution can also reduce the burden on referral care centres. However, its implementation feasibility and performance on diverse cameras in primary and secondary care settings need to be assessed in future research.

CRediT authorship contribution statement

Rohan Chawla: Writing – original draft, Resources, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Prachi Karkhanis:** Writing – original draft, Conceptualization. **Malay Shah:** Writing – original draft, Conceptualization. **Aritra Das:**

Writing – original draft, Validation. **Rishabh Sharma:** Software, Data curation. **Dhwani Almaula:** Conceptualization. **Pradeep Venkatesh:** Writing – review & editing, Validation, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Harsh Vardhan Singh:** Software. **Mukul Kumar:** Software. **Ramanuj Samanta:** Validation. **Vinod Kumarl:** Data curation. **Amar Shah:** Writing – review & editing. **Bhavin Vadera:** Writing – review & editing. **Nakul Jain:** Writing – review & editing. **Akanksha Sen:** Writing – review & editing. **Shyamsundar Shreedhar:** Writing – review & editing. **Vipin Garg:** Writing – review & editing. **Soma Dhaval:** Writing – review & editing. **Kowshik Ganesh:** Writing – review & editing. **Srinivas Rana:** Writing – review & editing, Software. **Radhika Tandon:** Writing – review & editing, Validation, Supervision, Resources, Methodology.

Declaration of competing interest

The AI solution discussed in this manuscript was developed as part of the United States Agency for International Development (USAID) funded project, with technical and AI/ML expertise led by Wadhwani AI which is the official AI partner with the Ministry of Health and Family Welfare (MoHFW) and has been operating as the eHealth AI unit for MoHFW since January 2022. To institutionalize the domain-specific development, MoHFW established Centers of Excellence (CoEs) for AI in Healthcare with All India Institute of Medical Sciences (AIIMS) Delhi leading the efforts. The evaluation of this AI solution was independently conducted at AIIMS Delhi, and did not receive funding from USAID. The content is solely the responsibility of the authors and does not necessarily represent the official views of USAID, AIIMS, or MoHFW. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

Authors express their sincere gratitude to the patients whose fundus images were used in the study, the doctors and healthcare workers at All India Institute of Medical Sciences, the officials at the Ministry of Health and Family Welfare, Government of India, and Wadhwani AI for their immense support in this study. We would also like to thank Dr. Shourya Vardhan Azad, Dr. Nawazish Fatima Sheikh, Dr. Mousumi Banerjee, Dr. Divya Agarwal, Dr. Pulak Agarwal, Dr. Sourabh Verma, Dr. SreeramJayaraj, Dr. Devesh Kumawat, Dr. Ayushi Sinha, Dr. Aishwarya Rathod, Dr. Aarush Deora, Dr. Debarun Sharma, Dr. Prabhav Puri, Dr. Vineet Batwani, Dr. Shashank Narang, Dr. Sindhuja Kandasamy, Dr. Deeksha Rani, Dr. Akshaya Balaji, Dr. Shakha Gupta, Dr. Somya Kumari, Dr. Ananya Kagilankar, and Dr. Anirudh Kapoor for their contributions in providing the labels for the fundus images.

Data availability

Data-sharing requests from individuals/research groups that submit a research proposal will be considered. Requests for de-identified data following manuscript publication should be directed to the corresponding author. Proposals will be evaluated internally, and the data will be shared solely based on scientific merit after signing a data-sharing agreement.

References

[1] Leasher JL, Bourne RRA, Flaxman SR, Jonas JB, Keeffe J, Naidoo K, et al. Global estimates on the number of people blind or visually impaired by diabetic retinopathy: a meta-analysis from 1990 to 2010. Diabetes Care 2016;39(9):1643–9.
[2] Raman R, Rani PK, Reddi Rachepalle S, Gnanamoorthy P, Uthra S, Kumaramanickavel G, et al. Prevalence of diabetic retinopathy in India: Sankara Nethralaya diabetic retinopathy epidemiology and molecular genetics study report 2. Ophthalmology 2009 Feb;116(2):311–8.
[3] Raman R, Ganesan S, Pal SS, Kulothungan V, Sharma T. Prevalence and risk factors for diabetic retinopathy in rural India. Sankara Nethralaya Diabetic Retinopathy Epidemiology and Molecular Genetic Study III (SN-DREAMS III), report no 2. BMJ Open Diabetes Res Care 2014;2(1):e000005.
[4] Raman R, Vasconcelos JC, Rajalakshmi R, Prevost AT, Ramasamy K, Mohan V, et al. Prevalence of diabetic retinopathy in India stratified by known and undiagnosed diabetes, urban–rural locations, and socioeconomic indices: results from the SMART India population-based cross-sectional screening study. Lancet Glob Health 2022;10(12):e1764–73.
[5] Yau JW, Rogers SL, Kawasaki R, Lamoureux EL, Kowalski JW, Bek T, et al. Global prevalence and major risk factors of diabetic retinopathy. Diabetes Care 2012 Mar;35(3):556–64.
[6] Takkar B, Das T, Thamarangsi T, Rani PK, Thapa R, Nayar PD, et al. Development of diabetic retinopathy screening guidelines in South-East Asia region using the context, challenges, and future technology. Semin Ophthalmol 2022;37(1):97–104. Jan 2.
[7] the IAPB South East Asia Region Eye Health Study Group, Das T, Ackland P, Correia M, Hanutsaha P, Mahipala P, et al. Is the 2015 eye care service delivery profile in Southeast Asia closer to universal eye health need! Int Ophthalmol. 2018 Apr;38(2):469–80.
[8] Sosale B, Sosale AR, Murthy H, Sengupta S, Naveenam M. Medios–an offline, smartphone-based artificial intelligence algorithm for the diagnosis of diabetic retinopathy. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC7003589/pdf/IJO-68-391.pdf.
[9] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7639):115–8. Feb 2.
[10] Titano JJ, Badgeley M, Schefflein J, Pain M, Su A, Cai M, et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. Nat Med 2018 Sep;24(9):1337–41.
[11] Fang L, Cunefare D, Wang C, Guymer RH, Li S, Farsiu S. Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search. Biomed Opt Express 2017;8(5):2732–44. May 1.
[12] Lee CS, Tyring AJ, Deruyter NP, Wu Y, Rokem A, Lee AY. Deep-learning based, automated segmentation of macular edema in optical coherence tomography. Biomed Opt Express 2017;8(7):3440–8. Jul 1.
[13] Schlegl T, Waldstein SM, Bogunovic H, Endstraßer F, Sadeghipour A, Philip AM, et al. Fully automated detection and quantification of macular fluid in OCT using deep learning. Ophthalmology 2018 Apr;125(4):549–58.
[14] Maloca PM, Lee AY, de Carvalho ER, Okada M, Fasler K, Leung I, et al. Validation of automated artificial intelligence segmentation of optical coherence tomography images. PloS One 2019;14(8):e0220063.
[15] Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316(22):2402. Dec 13.
[16] Grzybowski A, Brona P, Lim G, Ruamviboonsuk P, Tan GSW, Abramoff M, et al. Artificial intelligence for diabetic retinopathy screening: a review. Eye 2020;34(3):451–60. Mar 1.
[17] Nakayama LF, Zago Ribeiro L, Novaes F, Miyawaki IA, Miyawaki AE, De Oliveira JAE, et al. Artificial intelligence for telemedicine diabetic retinopathy screening: a review. Ann Med 2023;55(2):2258149. Dec 12.
[18] Rajesh AE, Davidson OQ, Lee CS, Lee AY. Artificial intelligence and diabetic retinopathy: AI framework, prospective studies, head-to-head validation, and cost-effectiveness. Diabetes Care 2023;46(10):1728–39.
[19] Ting DSW, Cheung CY, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. Jama 2017;318(22):2211–23. Dec 12.
[20] Huang X, Wang H, She C, Feng J, Liu X, Hu X, et al. Artificial intelligence promotes the diagnosis and screening of diabetic retinopathy. Front Endocrinol 2022;13:946915. Sep 29.
[21] Tan TE, Wong TY. Diabetic retinopathy: looking forward to 2030. Front Endocrinol 2023;13:1077669. Jan 9.
[22] ICO guidelines for diabetic eye care [Internet]. International Council of Ophthalmology.; [cited 2024 Dec 4]. Available from: https://icoph.org/eye-care-delivery/diabetic-eye-care/.
[23] Dugas E, Jorge J, Cukierski W. Diabetic retinopathy detection dataset [Internet]. San Francisco, CA, USA: Kaggle; 2015. Available from: https://www.kaggle.com/competitions/diabetic-retinopathy-detection.
[24] Karthik M, Dane S. APTOS 2019 blindness detection dataset [Internet]. Atlanta, GA, USA: Kaggle; 2019. Available from: https://www.kaggle.com/competitions/aptos2019-blindness-detection.
[25] Ocular disease recognition dataset [Internet]. Kaggle; 2020. Available from: https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k.
[26] Ramachandran R. The impact of artificial intelligence in screening for diabetic retinopathy in India. Available from: https://pmc.ncbi.nlm.nih.gov/articles/pmid/31827270/.
[27] Cohen Jérémie F, Korevaar Daniël A, Altman Douglas G, Bruns David E, Gatsonis Constantine A, Hooft Lotty, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open 2016;6(11):e012799.
[28] Farrington CP, Manning G. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. Stat Med 1990 Dec;9(12):1447–54.
[29] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159–74.
[30] Hripcsak G, Heitjan DF. Measuring agreement in medical informatics reliability studies. J Biomed Inform 2002 Apr;35(2):99–110.

[31] Raman R, Ramasamy K, Rajalakshmi R, Sivaprasad S, Natarajan S. Diabetic retinopathy screening guidelines in India: All India Ophthalmological Society diabetic retinopathy task force and Vitreoretinal Society of India Consensus Statement. Indian J Ophthalmol 2021 Mar;69(3):678–88.
[32] Daskivich LP, Vasquez C, Martinez C, Tseng CH, Mangione CM. Implementation and evaluation of a large-scale teleretinal diabetic retinopathy screening program in the Los Angeles County Department of Health Services. JAMA Intern Med 2017; 177(5):642–9. May 1.
[33] Ipp E, Liljenquist D, Bode B, Shah VN, Silverstein S, Regillo CD, et al. Pivotal evaluation of an artificial intelligence system for autonomous detection of referrable and vision-threatening diabetic retinopathy. JAMA Netw Open 2021;4 (11):e2134254. Nov 1.
[34] Padhy S, Takkar B, Chawla R, Kumar A. Artificial intelligence in diabetic retinopathy: a natural step to the future. Indian J Ophthalmol 2019;67(7):1004.
[35] Bhaskaranand M, Ramachandra C, Bhat S, Cuadros J, Nittala MG, Sadda SR, et al. The value of automated diabetic retinopathy screening with the EyeArt system: a study of more than 100,000 consecutive encounters from people with diabetes. Diabetes Technol Ther 2019;21(11):635–43. Nov 1.