Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Explainable and Trustworthy AI for Diabetic Retinopathy: A Systematic Review and Diagnostic Meta-analysis

Sunitha T

sunithat.mca@bmsce.ac.in

B.M.S. College of Engineering

Ram Mohan Reddy Ch

B.M.S. College of Engineering

Article

Keywords: Diabetic retinopathy, Explainable artificial intelligence, Deep learning, Diagnostic meta-analysis, Trustworthy AI, Clinical Validation

Posted Date: October 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10997759/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Although Deep Learning (DL) is increasingly being used for Diabetic Retinopathy (DR) detection and severity grading, there is a need for more than diagnostic accuracy to translate into the clinic. This systematic review and diagnostic meta-analysis assessed the state-of-the-art explainable and trustworthy-AI methods for DR, combining diagnostic performance with interpretability, calibration, uncertainty, generalizability, fairness, privacy, and deployment evidence. Based on PRISMA 2020 and PRISMA-DTA guidelines, a search for eligible studies published between January 2022 and August 2026 was conducted on PubMed/MEDLINE, Web of Science, IEEE Xplore, and ScienceDirect. Ten studies met the final inclusion criteria and six studies provided complete or unequivocally reconstructable 2 × 2 data for clinically compatible Any-DR versus No-DR endpoints. A Reitsma-style bivariate random-effects model with restricted maximum-likelihood estimation was used to jointly synthesize sensitivity and false-positive rate, and SROC, prediction-region, leave-one-out and exploratory Deeks' analyses were also conducted. The pooled sensitivity was 0.9842 and the specificity was 0.9760, with high average discrimination and a wide prediction region, reflecting uncertainty in future clinical settings. Generally, binary DR detection demonstrated more consistent diagnostic performance than multiclass severity grading. The use of explainability methods such as Grad-CAM, SHAP, attention visualization and interpretable evidence maps was common, while quantitative or clinician-based validation was still rare. External validation was inconsistent and calibration, subgroup fairness, privacy-preserving learning and prospective human-AI evaluation were limited. Multicentre validation, standardized endpoints, validated explanations, calibrated uncertainty, fairness assessment, privacy safeguards, accountable human oversight, and post-deployment monitoring, across diverse populations, devices, and clinical workflows and settings are therefore required for trustworthy clinical deployment.

1. Introduction

1.1 Background and Clinical Importance of Diabetic Retinopathy

Diabetic Retinopathy (DR) is one of the most important microvascular complications of diabetes and a leading cause of preventable vision loss and blindness. Hyperglycaemia continues to cause progressive damage to the retinal microvasculature, leading to the development of microaneurysms, haemorrhages, exudates, vascular abnormalities and, in late disease, to the development of irreversible changes in the retina that lead to permanent visual loss. Early-stage DR is often without symptoms and systematic retinal screening is key to early diagnosis, referral and treatment [1]. The current standard of DR screening is mainly based on fundus imaging assessed by ophthalmologists or trained graders. While clinically proven, manual assessment is resource intensive, prone to interobserver variability and not easily scalable in situations with limited specialist availability. Automated retinal-image analysis thus offers a valuable opportunity to increase screening capacity, especially when used with portable or non-mydriatic imaging systems in resource-poor settings [2]. But overall accuracy is not enough to determine clinical utility. Screening systems must be highly sensitive to reduce the number of missed disease cases and have an acceptable specificity to reduce the number of unnecessary referrals. Performance may also be influenced by the prevalence of disease, image quality, imaging device, population characteristics and clinical setting. Artificial intelligence (AI) for DR must therefore be clinically useful, meaning it must have diagnostic discrimination, be robust, have external validity, be interpretable, and operate reliably across the intended deployment environments.

1.2 Deep Learning, Explainability, and Trustworthy AI for DR

Retinal-image analysis has progressed from traditional machine-learning methods that rely on handcrafted features to deep-learning methods that learn complex hierarchical representations directly from fundus images. Convolutional neural networks (CNNs) continue to be the popular choice for DR detection and severity classification and transfer learning allows for the transfer of pre-trained networks to ophthalmic data with relatively small task-specific datasets. Retinal lesion information has been integrated into transfer-learning pipelines, which has led to improved disease discrimination [3] and validation studies have shown their applicability across DR severity categories [4]. New advancements include ensemble learning, attention mechanisms, hybrid CNN-transformer architectures, and Vision Transformers. Ensemble models have

been used to combine complementary feature representations and EfficientNet-based systems have been shown to have good diagnostic capability [5]. ResViT [6] is a hybrid method that combines convolutional feature extraction and global representation learning using attention. The use of DR-AI applications has also progressed beyond binary screening to multiclass severity grading and longitudinal prediction of progression, which could allow for risk-stratified screening intervals and personalised follow-up [7]. As architectures grow more complex, however, there has been growing concern about model opacity. To find out which parts of the image are important to the prediction, explainable artificial intelligence (XAI) techniques like Gradient-weighted Class Activation Mapping (Grad-CAM), SHAP, attention visualization, and other saliency methods are now being used more often [8, 9]. However, visually plausible explanations do not necessarily constitute clinically valid model reasoning.

Correspondence between model explanations and recognised pathological features, expert-defined annotations or lesion-level reference standards are required for stronger evidence. Interactive retinal diagnostic systems are starting to combine the results with disease prediction [10] and tools like MAPLES-DR offer detailed anatomical and pathological annotations to support objective assessment of the quality of explanations [11]. Explainability is one piece of the puzzle of trustworthy AI. Clinically reliable systems should also consider the problem of predictive uncertainty, calibration, robustness, fairness, external generalizability, privacy, and accountable human oversight. Low confidence predictions can be identified by Bayesian and uncertainty-aware methods for specialist review [12]. Fairness assessment is also crucial, as diagnostic errors can vary between demographic groups, disease stages, imaging devices, or acquisition environments if datasets are not sufficiently representative [13]. The importance of privacy-preserving approaches is growing, as high-quality retinal datasets may be held in different institutions and there may be restrictions on centralized data sharing. Clinical governance also differentiates experimental performance from deployable medical AI. An independent diabetic-retinopathy detection device has been authorized before by the US Food and Drug Administration, showing that AI for the retina can move beyond research in decision support into regulated autonomous screening.

On a governance level, the World Health Organization (WHO) [35] highlights the need for protection of autonomy, safety, transparency, accountability, inclusiveness, equity, and ongoing evaluation throughout the life cycle of AI systems in health. Medical-device AI could be considered in the high-risk category of the EU AI Act if the AI system itself is a medical device or safety component subject to a third-party conformity assessment [36], thereby connecting AI governance with current medical-device regulation. The regulatory developments further highlight the importance of assessing DR-AI based on clinical evidence, risk management, transparency, human oversight, and post-market monitoring, not just on accuracy. Autonomous DR screening shows evidence of AI's ability to enhance screening and follow-up in care pathways [14]. Clinically deployable DR-AI should therefore be reliable across populations and imaging settings, give calibrated confidence estimates, give clinically meaningful explanations, show equitable performance across sub-populations, protect sensitive data, and allow for appropriate human oversight. Assessment should thus include workflow compatibility, patient-level outcomes, human-AI interaction, and post-deployment surveillance.

1.3 Research Gap, Aim, and Objectives

Previous DR-AI studies have mainly concentrated on the model architecture, preprocessing methods, benchmark datasets, and overall classification accuracy. This is often reported with high accuracy, but this can mask clinically important false negative and false positive errors. Sensitivity and specificity are especially important in the context of screening, and previous diagnostic meta-analyses have shown the benefit of simultaneously synthesising these measures to evaluate AI-based DR screening [15]. There is still a large gap, however, between technical performance and evidence for trustworthy clinical translation. Models with good performance in one population can suffer from a loss of accuracy in another population, or when tested in another imaging device, or when the distribution changes. External validation is not consistent and quantitative or clinician-based validation of XAI outputs is rare. Formal evidence of subgroup fairness, privacy protection, and future human-AI workflow performance is scarce, and calibration and uncertainty are rarely systematically evaluated [34].

Confidence in generalizability is also limited by dataset imbalance, poor reporting of demographics, device heterogeneity, repeated use of public benchmark datasets, and lack of prospective validation. Most of the previous reviews have focused on diagnostic performance, model development or explainability individually. Diagnostic-accuracy reviews have focused on summarising screening performance, while XAI studies have focused mainly on visualization approaches. A crucial question yet to be answered is whether high diagnostic performance is accompanied by the broader characteristics required for trustworthy clinical AI. There is limited integrated evidence that brings together validated explainability, calibration and uncertainty, external generalizability, fairness, privacy, human-AI interaction and deployment readiness. In this systematic review and diagnostic meta-analysis, the authors assess the current deep-learning systems for DR from the diagnostic-performance and trustworthy-AI perspectives. It integrates quantitative synthesis of sensitivity and specificity with structured assessment of explainability, calibration, uncertainty, external validation, fairness, privacy, human-AI interaction, and clinical deployment evidence.

The objectives were to:

1. Characterize contemporary deep-learning architectures, datasets, validation strategies, and diagnostic tasks for DR detection and severity grading.

2. Synthesize diagnostic performance and meta-analyse compatible 2 × 2 data using sensitivity and specificity.

3. Evaluate XAI approaches and determine whether explanations have been quantitatively or clinically validated.

4. Assess evidence relating to uncertainty, calibration, fairness, external generalizability, privacy preservation, human-AI interaction, and readiness for clinical deployment.

2. Materials and Methods

2.1 Review Design and Reporting

To assess deep-learning methods for diabetic retinopathy (DR) detection and severity grading, a systematic review with diagnostic-accuracy meta-analysis was performed, focusing on explainable and trustworthy artificial intelligence. Reporting was done in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement [16] and the diagnostic-accuracy components were organised following the PRISMA-DTA [38] principles. The review combined quantitative diagnostic synthesis and qualitative structured analysis of explainability and trustworthy-AI features. Meta-analysis was conducted on studies that reported fully or unambiguously reconstructable 2 × 2 diagnostic data for clinically similar binary DR outcomes. Studies that did not meet the criteria for quantitative pooling were kept for descriptive synthesis if they provided relevant evidence on model performance, explainability, calibration, external validation, fairness, privacy, human-AI interaction, or clinical deployment.

PRISMA-DTA [38] (Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy Studies) extends the PRISMA framework for reviews evaluating diagnostic tests. Unlike conventional systematic reviews, diagnostic-accuracy reviews require transparent reporting of index tests, reference standards, target conditions, diagnostic endpoints, and methods used to synthesize sensitivity and specificity. Therefore, PRISMA-DTA [38] was adopted to improve methodological transparency, reproducibility, and reporting quality for the diagnostic meta-analysis component of this review.

2.2 Protocol and Review Framework

Eligibility criteria, data-extraction domains, and the primary diagnostic outcome were defined before final study inclusion. Some details of the quantitative synthesis and trustworthy-AI categorization were refined during review development. Because the review protocol was not prospectively registered, the possibility of post-hoc methodological modification cannot be completely excluded and is acknowledged as a limitation. Studies were assessed using a Population, Index Test, Reference Standard, and Diagnostic Target Condition (PIRD) framework. Diagnostic-performance and trustworthy-AI

evidence were treated as methodologically distinct but complementary domains. Accordingly, external validation, reader evaluation, federated learning, prototype interfaces, and prospective clinical deployment were assessed separately and were not considered interchangeable indicators of clinical readiness.

The review protocol was not prospectively registered in PROSPERO or another systematic-review registry. Therefore, the possibility of post-hoc methodological modification cannot be completely excluded. The absence of prospective registration is acknowledged as a methodological limitation and should be considered when interpreting the reproducibility of the review.

2.3 PIRD Framework and Eligibility Criteria

Eligible Index Tests were deep-learning-based diagnostic systems for DR detection, screening, or severity grading, such as convolutional neural networks, transfer-learning architectures, ensemble models, attention-based networks, Vision Transformers, hybrid deep-learning models, and other deep-learning approaches for DR detection, screening, or severity grading.

Criterion	Inclusion	Exclusion
Population/data	Patients or clinically relevant retinal images evaluated for DR	Non-DR populations or irrelevant imaging data
Index test	Substantial deep-learning component	Conventional ML without substantive DL
Target	DR detection, screening, or severity grading	Unrelated ophthalmic conditions
Reference standard	Clinical/expert grading or documented ophthalmic dataset labels	Inadequately defined reference labels
Publication	Peer-reviewed primary research	Reviews, meta-analyses, editorials, commentaries
Period	January 2022-August 2026	Outside predefined period
Language	English	Non-English
Meta-analysis	Compatible Any-DR vs No-DR 2 × 2 data	Incompatible endpoint or unreconstructable binary data

Studies excluded from quantitative synthesis could remain eligible for the systematic review when they contributed evidence relevant to explainability or trustworthy AI.

2.4 Information Sources and Search Strategy

PubMed/MEDLINE, Web of Science, IEEE Xplore, and ScienceDirect were searched for peer-reviewed studies published from January 2022 to August 2026. Search concepts covered DR, deep learning, diagnostic accuracy, explainability, and trustworthy-AI characteristics.

Core terms included “diabetic retinopathy”, “deep learning”, “convolutional neural network”, “CNN”, “Vision Transformer”, “artificial intelligence”, “diagnostic accuracy”, “sensitivity”, “specificity”, “explainable artificial intelligence”, “XAI”, “Grad-CAM”, “SHAP”, “uncertainty”, “calibration”, “fairness”, “privacy”, and “clinical deployment”. Boolean operators AND and OR were used to combine concept groups, with syntax adapted to each database interface.

The conceptual search structure was:

("diabetic retinopathy") AND ("deep learning" OR CNN OR "convolutional neural network" OR "Vision Transformer" OR "artificial intelligence") AND ("diagnostic accuracy" OR sensitivity OR specificity OR "explainable AI" OR XAI OR Grad-CAM OR SHAP OR uncertainty OR calibration OR fairness OR privacy OR deployment).

References of potentially eligible articles and relevant reviews were also manually searched. Only English language peer-reviewed literature was searched in the defined publication period. Complete database-specific search strategies, search fields, and applied limits are provided in Supplementary Table S1. The final search was performed on 2 August 2026.

2.5 Study Selection

Citations were deduplicated and consolidated prior to screening. Titles and abstracts were evaluated using predefined eligibility criteria and full-text assessment of potentially relevant reports was conducted. Reasons for exclusion were recorded and summarised in the PRISMA flow diagram. Screening always took into account the inclusion of a significant deep-learning component, the relevance to DR detection or grading, the quality of the reference standard, the availability of interpretable diagnostic output, and the contribution to predefined trustworthy-AI domains. Study selection was conducted in accordance with the PRISMA 2020 guidelines. A total of 268 records were identified through database searching (PubMed/MEDLINE, Web of Science, IEEE Xplore, and ScienceDirect; n = 248) and other sources, including manual reference checking (n = 20). After removal of 52 duplicate records, 216 records underwent title/abstract screening, of which 178 were excluded as not relevant. Title/abstract and full-text screening were independently performed by both authors. Disagreements were resolved through discussions and consensus. Of the remaining 38 reports sought for retrieval, 3 could not be obtained, leaving 35 full-text reports assessed for eligibility. Of these, 25 were excluded for the following reasons: irrelevance to diabetic retinopathy (n = 7), no substantial deep-learning component (n = 6), insufficient diagnostic-accuracy data (n = 5), no predefined XAI/trustworthy-AI outcome (n = 4), and review/editorial articles (n = 3). This resulted in 10 studies included in the systematic review, of which 6 provided sufficient data for inclusion in the diagnostic meta-analysis. Data extraction performed by Sunitha T using a structured extraction framework and independently verified by Ch. Ram Mohan Reddy. Discrepancies were resolved by consensus.

2.6 Data Extraction and Endpoint Harmonization

Study characteristics, datasets, model architecture, diagnostic design, validation strategy, performance measures, and trustworthy-AI evidence were captured using a structured extraction framework. Variables extracted were publication year, dataset and sample size, imaging modality, diagnostic endpoint, reference standard, preprocessing, class-balancing strategy, architecture, validation design, external validation and, if available, dataset origin, imaging device, and patient-level versus image-level partitioning.

The diagnostic results were accuracy, sensitivity, specificity, AUC, precision, F1-score, TP, FN, FP, TN. For studies that provided complete 2 × 2 data, likelihood ratios and diagnostic odds ratios were calculated. Confusion matrix was analyzed for multiclass studies to see if an unambiguous binary endpoint could be generated. Only when the class mapping was explicit and reproducible was reconstruction performed, with grade 0 being classified as No-DR, and all grades indicating detectable retinopathy being classified as Any-DR. Direct and reconstructed 2 × 2 data were identified separately to ensure audit trail. Trustworthy-AI extraction comprised XAI technique, explanation-validation approach, lesion correspondence, clinician/reader evaluation, uncertainty, calibration, external validation, fairness, privacy-preserving methodology, human-AI interaction, and deployment evidence.

2.7 Risk-of-Bias and Applicability Assessment

The QUADAS-2 [37] was used to evaluate risk of bias in Patient Selection, Index Test, Reference Standard, and Flow and Timing. Judgments were categorised as low, high or unclear risk, with unclear risk given when reporting was inadequate to make a judgement. Sampling and spectrum bias, independence of the index test and reference standard, adequacy of the reference standard, and participant/image flow were considered in the assessment. Applicability concerns were addressed regarding the population, index test, and target condition included in the study being relevant to the review question.

Since QUADAS-2 [37] was not specifically designed for AI, other methodological issues were also explored qualitatively, such as data leakage, class imbalance, train-test independence, repeated use of benchmark-datasets, demographic reporting, domain shift, calibration, external validation and explanation validity. Study-level QUADAS-2 judgments are provided in

Supplementary Table S4. The signalling questions and judgment criteria used for the four QUADAS-2 domains are provided in Supplementary Table S5.

2.8 Diagnostic-Accuracy Meta-analysis

Only studies that reported 2 × 2 data that were complete or unequivocally reconstructable for clinically compatible Any-DR versus No-DR outcomes were included in the quantitative synthesis. Study-specific sensitivity and specificity were calculated as:

$$Sensitivity = \frac{TP}{TP+FN} \quad Specificity = \frac{TN}{TN+FP}$$

Exact binomial 95% confidence intervals were calculated for study-level estimates.

Sensitivity and false-positive rate (FPR = 1 − specificity) were jointly synthesized on the logit scale using a Reitsma-style bivariate random-effects model:

$$y_i = \begin{pmatrix} \log\left(TP_i/FN_i\right) \\ \log\left(FP_i/TN_i\right) \end{pmatrix}$$

Within-study variances were approximated as:

$$Var\left(\text{logit Se}\right) = \frac{1}{TP} + \frac{1}{FN} \quad Var\left(\text{logit FPR}\right) = \frac{1}{FP} + \frac{1}{TN}$$

The between-study heterogeneity was modelled by an unstructured 2 × 2 covariance matrix with separate random-effects variances and the correlation between them. Restricted maximum likelihood was used to estimate the parameters. Standard deviations between studies were parameterized on log scale and correlation was constrained by hyperbolic-tangent transformation. Numerical optimization was performed with L-BFGS-B with bounded parameters, multiple starting values and explicit convergence checking. A prespecified continuity correction of 0.5 was available for zero diagnostic cells; however, no correction was required because all four cells were non-zero in each included study.

No correction was required when all four cells were non-zero. Summary sensitivity and FPR were back-transformed to the probability scale and specificity was calculated as (1 − FPR). Logit scale 95 per cent confidence intervals were calculated and back transformed. Heterogeneity was assessed mainly using between-study variance components, transformed-outcome correlation, dispersion of study operating points, and the SROC prediction region, and not the traditional univariate I^2 or Cochran Q statistics.

2.9 SROC and Sensitivity Analyses

A model-based summary ROC curve was obtained by applying a Rutter-Gatsonis-type transformation [40] to the fitted bivariate model [41]. A summary operating point, 95% confidence region, and 95% prediction region were developed. The confidence region was used to capture the uncertainty about the mean operating point, while the prediction region was used to capture the variability in a future similar study. The robustness of the results was evaluated by leave-one-out analysis, refitting the full bivariate model with one study removed at a time. Sensitivity, specificity, confidence intervals and covariance parameters were re-estimated after each exclusion. Formal subgroup meta-analysis was not performed because of the small number of quantitatively comparable studies, which would have resulted in unstable subgroup estimates.

2.10 Small-Study Effects and Statistical Implementation

The quantitative analyses were carried out using Python 3.10.11. The data was processed and numerical operations were carried out using NumPy 1.26.4 and pandas 2.3.3. The statistical optimization, probability transformations, exact confidence-interval calculations, and other numerical procedures were done using SciPy 1.15.3, and the proportion-based calculations and weighted regression were done using statsmodels 0.14.6. Matplotlib 3.10.8 was used to generate graphical outputs like forest plots, SROC visualization and funnel plots. The implementation of the bivariate diagnostic model was done using

scipy.optimize.minimize with the L-BFGS-B optimization algorithm. Confidence and prediction regions were formed from the estimated covariance matrices with the help of the chi-square distribution with 2 degrees of freedom.

2.11 Explainability and Trustworthy-AI Synthesis

The XAI evidence was categorized as qualitative visualization, quantitative lesion/localization validation, clinician-based validation, or structured reader/human-AI evaluation. Trustworthy-AI evidence was evaluated for uncertainty/calibration, external generalizability, fairness/representativeness, privacy/security, human-AI interaction, and clinical deployment. Each domain was classified as clearly assessed, partially assessed, or not substantially assessed.

3. Results

3.1 Study Selection

A total of 268 records were identified from the systematic search (248 records from PubMed/MEDLINE, Web of Science, IEEE Xplore and ScienceDirect and 20 records from manual reference checking). 216 unique records were identified after removing 52 duplicates and were screened based on title and abstract. Of these, 178 were excluded and 38 reports were sought for full-text retrieval. Three reports were not retrieved, and 35 studies were assessed for full text.

Twenty-five were excluded due to being unrelated to DR, not having a significant deep-learning component, lacking relevant information on diagnostic performance, failing to address a predefined explainability/trustworthy-AI domain, or being review/editorial literature. Therefore, 10 main studies were selected for systematic review with 6 of them having complete or unequivocally reconstructable 2 × 2 data for diagnostic meta-analysis. While the eligibility period was from 2022 to 2026, all studies included in the final eligibility criteria were published in 2025–2026, suggesting that the evidence base is very contemporary but also very timebound. The study selection process following the PRISMA 2020 guidelines is illustrated in Fig. 1.

3.2 Study, Dataset, Model, and Validation Characteristics

The 10 studies included assessed heterogeneous deep-learning architectures, diagnostic endpoints, datasets, and validation strategies. The most common imaging modality was colour fundus photography. Common public datasets were APTOS 2019, EyePACS, IDRiD, DDR, Messidor-2 and SUSTech-SYSU. Mutawa et al. [20] also created a local database of 806 retinal images, which were graded by two ophthalmologists with adjudication by a senior retinal specialist.

The architectures included transformer ensembles, conventional CNNs, attention enhanced residual networks, YOLO systems, EfficientNet/ConvNeXt ensembles, and federated MobileNet architectures. Rahaman et al. [17] combined ViT, DeiT, and BEiT for binary DR classification. Djoumessi et al. [18] created an inherently interpretable Sparse BagNet and tested it on 10 external datasets. Ashraf et al. [19] evaluated HIRD-Net on multiple public datasets, and Ghosh et al. [22] used tri-attention within a ResNet-based system and validated it on IDRiD, which is an external dataset. Herrero-Tudela et al. [24] conducted one of the most extensive cross-dataset analyses of five publicly available datasets. The characteristics of the included deep-learning studies, including datasets, model architectures, diagnostic tasks, validation strategies, and explainability methods, are summarized in Table 1.

Table 1
Characteristics of the included deep-learning studies

Study	Dataset/sample	Model	Task	Validation	XAI
Rahaman et al. [17]	APTOS; n = 3,661	ViT + DeiT + BEiT	Any DR vs No DR	Internal	Attention/CAM
Djoumessi et al. [18]	EyePACS; large development cohort	Sparse BagNet	Early/Any DR	10 external datasets	Evidence maps
Ashraf et al. [19]	IDRiD/APTOS, DDR, EyePACS	HIRD-Net + attention	Binary + five-grade	Cross-dataset	Grad-CAM
Mutawa et al. [20]	Local; n = 806	YOLOv8/YOLO11	Binary + five-class	Patient-level	Eigen-CAM
Mahipal & Kumar [21]	APTOS; TJDR n = 236	ResNet50/DenseNet121/InceptionV3	Five-class	Held-out/XAI validation	Multiple CAM methods
Ghosh et al. [22]	APTOS + IDRiD	ResNet50 + tri-attention	Four-class	Internal + external	Grad-CAM
Civit-Masot et al. [23]	Messidor-2; n = 1,788	EfficientNetV2 + ConvNeXt	Five-class/referable DR	Cross-validation	Configurable heatmaps
Herrero-Tudela et al. [24]	Five public datasets	ResNet50	Five-class	Cross-dataset	SHAP
Ghadai et al. [25]	APTOS	EfficientNet-B1	Five-class	Held-out + 5-fold CV	Grad-CAM
Ciaramella et al. [26]	APTOS; n = 3,662	Federated MobileNet	Any DR vs No DR	20 non-IID clients	Grad-CAM variants

Repeated use of APTOS and related benchmark datasets indicates that the 10 studies should not be interpreted as 10 completely independent clinical populations.

3.3 Diagnostic Performance of Individual Models

The diagnostic performance was dependent on endpoint, class structure, architecture and validation design. Generally, binary DR detection yielded better discrimination than multiclass severity grading.

Rahaman et al. [17] reported accuracy of 0.9754, sensitivity of 0.9866, F1-score of 0.9761, and AUC of 0.9967. Ashraf et al. [19] obtained a binary accuracy of 99.50%, sensitivity of 99.7% and specificity of 99.3%. Mutawa et al. [20] reported binary accuracy of 97.02%, recall of 97.01%, and AUC of 0.98. Ciaramella et al. [26] tested the privacy-preserving federated model and found a lower sensitivity of 0.791 and derived specificity of 0.834.

Multiclass grading was more variable. Mahipal and Kumar [21] obtained 85.29% accuracy and QWK of 0.9295, and Ghadai et al. [25] obtained 84.97% accuracy and macro-F1 of 0.6627 with test-time augmentation. Herrero-Tudela et al. [24] reported an accuracy of 94.64%, sensitivity of 86.61%, specificity of 96.98%, and AUC of 0.98 for APTOS. The principal diagnostic performance metrics reported across the included studies are presented in Table 2.

Table 2
Principal diagnostic performance of included studies

Study	Endpoint	Accuracy	Sensitivity	Specificity	AUC
Rahaman et al. [17]	Any DR vs No DR	0.9754	0.9866	0.9640*	0.9967
Djoumessi et al. [18]	Early/Any DR	0.9060	0.7090	0.9770	0.9040
Ashraf et al. [19]	Binary screening	0.9950	0.9970	0.9930	NR
Mutawa et al. [20]	Binary DR	0.9702	0.9701	NR	0.9800
Mahipal & Kumar [21]	Five-class DR	0.8529	NR	NR	NR
Ghosh et al. [22]	Four-class DR	0.9701/0.9940	0.9826/0.9920	NR	0.9853/0.9944
Civit-Masot et al. [23]	Five-class/referable DR	0.9670	0.9800	0.9600	≈ 0.9800
Herrero-Tudela et al. [24]	Five-class DR	0.9464	0.8661	0.9698	0.9800
Ghadai et al. [25]	Five-class + TTA	0.8497	Class-dependent	Class-dependent	0.9040
Ciaramella et al. [26]	Any DR vs No DR	≈ 0.812	0.791	0.834*	0.812

*Derived from available confusion-matrix counts. NR, not reported.

3.4 Diagnostic Meta-analysis

Six studies provided directly reported or reconstructable 2 × 2 data for clinically compatible Any-DR versus No-DR endpoints. Rahaman et al. [17], Ashraf et al. [19], and Ciaramella et al. [26] contributed direct binary data, whereas Mahipal and Kumar [21], Ghosh et al. [22], and Ghadai et al. [25] were reconstructed from published multiclass confusion matrices. The reconstructed and directly reported diagnostic 2 × 2 contingency data used for the bivariate meta-analysis are provided in Table 3.

Table 3
Diagnostic 2 × 2 data included in the bivariate meta-analysis

Study	Endpoint	TP	FN	TN	FP	Source
Rahaman et al. [17]	Any DR vs No DR	367	5	348	13	Direct
Ashraf et al. [19]	Abnormal vs normal	2,198	7	1,959	14	Direct
Mahipal & Kumar [21]	Derived Any DR	183	3	179	2	Reconstructed
Ghosh et al. [22]	Derived Any DR, IDRiD	373	1	123	1	Reconstructed
Ghadai et al. [25]	Derived Any DR	163	4	195	4	Reconstructed
Ciaramella et al. [26]	Any DR vs No DR	148	39	151	30	Direct

The bivariate random-effects model yielded pooled sensitivity of **0.9842** and pooled specificity of **0.9760**. Five studies clustered within a high-sensitivity/high-specificity region, whereas Ciaramella et al. [26] showed substantially lower paired performance. These estimates demonstrate high average discrimination for quantitatively compatible binary DR tasks but should not be interpreted as universal clinical operating characteristics because of between-study variation. The pooled sensitivity and specificity estimates, together with study-specific confidence intervals, are presented in Fig. 2.

3.5 SROC and Sensitivity Analysis

The bivariate model produced a summary operating point within the high-discrimination region of ROC space, with a model-derived SROC AUC of approximately **0.9941**. The 95% prediction region was substantially wider than the 95% confidence region, indicating greater uncertainty regarding diagnostic performance in a future independent clinical population than around the estimated average operating point. The summary receiver operating characteristic (SROC) curve and the corresponding confidence and prediction regions are shown in Fig. 3.

Leave-one-out analysis indicated that the overall interpretation was not materially altered by exclusion of individual studies. Pooled sensitivity ranged from **0.9780 to 0.9911**, while specificity ranged from **0.9682 to 0.9840** after sequential study exclusion. Removing Ciaramella et al. [26] produced the highest pooled sensitivity and specificity, confirming that this study influenced the magnitude, but not the overall interpretation, of diagnostic discrimination.

3.6 Small-Study Effects and Risk of Bias

Deeks' funnel-plot [39] asymmetry test yielded **p = 0.1077**. Because only six studies were available, the analysis was considered exploratory and underpowered; therefore, the non-significant result was not interpreted as evidence that publication bias was absent. Potential small-study effects were explored using the Deeks' funnel plot shown in Fig. 4.

QUADAS-2 [37] assessment showed that Patient Selection and Reference Standard were the domains most frequently assigned unclear judgments. Index Test was rated low risk across all studies, and Flow and Timing was low risk in nine studies. Ashraf et al. [19] was judged high overall risk because of concerns in Patient Selection and Flow and Timing. The study-level QUADAS-2 risk-of-bias assessment is summarized in Table 4.

Table 4
QUADAS-2 risk-of-bias assessment

Study	Patient selection	Index test	Reference standard	Flow/timing	Overall
Rahaman [17]	Unclear	Low	Unclear	Low	Unclear
Djoumessi [18]	Unclear	Low	Low	Low	Unclear
Ashraf [19]	High	Low	Unclear	High	High
Mutawa [20]	Unclear	Low	Low	Low	Unclear
Mahipal [21]	Unclear	Low	Unclear	Low	Unclear
Ghosh [22]	Low	Low	Unclear	Low	Unclear
Civit-Masot [23]	Unclear	Low	Low	Low	Unclear
Herrero-Tudela [24]	Unclear	Low	Unclear	Low	Unclear
Ghadai [25]	Unclear	Low	Unclear	Low	Unclear
Ciaramella [26]	Unclear	Low	Unclear	Low	Unclear

The study-level QUADAS-2 traffic-light visualization is presented in Fig. 5A. The distribution of QUADAS-2 judgments across the four risk-of-bias domains is illustrated in Fig. 5B.

3.7 Explainability and Clinical Interpretability

Explainability was present in all the included studies, with objective validation being rare. Several studies used Grad-CAM or related CAM approaches [19, 22, 23, 25, 26], attention-based visualization was used in [17], Eigen-CAM was used in [20], SHAP was used in [24] and evidence maps were used in [18] which are inherently interpretable.

Djoumessi et al. [18] reported a lesion-localization precision of 0.960 (95% CI 0.941–0.976) and also performed a retrospective reader study. Mahipal and Kumar [21] quantitatively evaluated the CAM output on 236 lesion annotated images

and achieved a maximum weighted lesion-overlap score of 0.721. The majority of the studies that remained gave qualitative explanations and were not lesion-level or clinician validated. The explainability methods and their corresponding validation maturity across the included studies are summarized in Table 5.

Table 5
Explainability methods and validation maturity

Study	XAI method	Quantitative validation	Human validation	Evidence level
Rahaman [17]	Attention/CAM	No	No	Qualitative
Djoumessi [18]	Evidence maps	Lesion precision 0.960	Reader study	Clinical/human
Ashraf [19]	Grad-CAM	No formal metric	No	Qualitative
Mutawa [20]	Eigen-CAM	No	Interface only	Qualitative
Mahipal [21]	Multiple CAM methods	Lesion overlap 0.721	No	Quantitative
Ghosh [22]	Grad-CAM + attention	No	No	Qualitative
Civit-Masot [23]	Configurable heatmaps	Partial	Expert-oriented	Intermediate
Herrero-Tudela [24]	SHAP	No lesion metric	No	Qualitative
Ghadai [25]	Grad-CAM	No	No	Qualitative
Ciaramella [26]	CAM variants	No	No	Qualitative

3.8 Trustworthy-AI Evidence

The most extensive external validation was reported by Djoumessi et al. [18], Herrero-Tudela et al. [24] and Ghosh et al. [22]. Calibration evidence was scant; Ghadai et al. [25] reported an expected calibration error of 0.0327 which dropped to 0.0239 after test-time augmentation. Civit-Masot et al. [23] added a confidence-quality factor for uncertain predictions.

Ciaramella et al. [26] specifically tackled privacy-preserving learning with federated learning over 20 non-IID clients. Structured human-AI evaluation was reported by Djoumessi et al. [18] through a retrospective reader study, whereas Mutawa et al. [20] reported a prototype Gradio interface without clinical deployment. There was no study that showed a comprehensive demonstration of multicentre prospective deployment, subgroup fairness, calibrated uncertainty, workflow effectiveness, and post-deployment surveillance.

3.9 Integrated Evidence

Overall, diagnostic discrimination and qualitative explainability were significantly more advanced than quantitative XAI validation, calibration, fairness assessment, privacy-preserving learning, structured human-AI evaluation, and prospective deployment. External validation, reader studies, and federated learning should not be used interchangeably as indicators of clinical readiness, but are important types of evidence. An integrated overview of explainability and trustworthy-AI evidence across the included studies is presented in Fig. 6.

4. Discussion

4.1 Principal Findings

This systematic review and diagnostic meta-analysis revealed substantial gaps in the evidence supporting trustworthy clinical translation of the diagnostic performance of deep-learning systems for Diabetic Retinopathy (DR). The pooled sensitivity and specificity for clinically compatible Any-DR and No-DR endpoints were 0.9842 and 0.9760, respectively, using the bivariate meta-analysis. However, the wide prediction region indicated substantial uncertainty in the performance expected in future clinical settings. Multi-class severity classification was not as consistent as binary detection, especially

for adjacent and under-represented stages. Explainability was present in all studies, but quantitative or clinician-based validation was not frequently used. External validation, calibration, uncertainty quantification, fairness, privacy-preserving learning, prospective human-AI evaluation, and post-deployment evidence were less mature than conventional performance reporting.

4.2 Diagnostic Performance and Generalizability

The pooled estimates suggest that deep learning can be useful for DR screening and triage, but should not be viewed as universally applicable clinical operating characteristics. Average discrimination does not imply equal performance across populations, devices, or healthcare settings. The data distribution and modelling context also affect the performance as shown by Ciaramella et al. [26] who reported lower performance. Multiclass grading is still more challenging due to the ordinal nature of DR, the overlapping visual features of DR grades, and the underrepresentation of advanced DR grades. Transportability can also be further decreased by variation in reference grading, preprocessing, image quality, acquisition devices and disease prevalence.

Architectural developments such as hybrid CNNs, attention mechanisms, and graph-based models [27–29] can help representation learning, but representation learning does not guarantee generalizability. In addition, the use of benchmark datasets such as APTOS, EyePACS, IDRiD, and others repeatedly, also restricts independence across the evidence base. Future studies should thus focus on multicentre external validation, heterogeneous imaging devices, strict patient level separation of development and test cohort, clinically relevant operating thresholds and evaluation under realistic acquisition conditions. Illumination, focus, resolution and field of view differences in image quality are still significant sources of degradation [30].

4.3 Explainability, Uncertainty, and Fairness

Visual plausibility is not sufficient to guarantee clinically valid reasoning, and Grad-CAM, CAM variants, SHAP, transformer attention, Eigen-CAM, and interpretable evidence maps were the most common. Objective XAI evaluation should be based on the correspondence with lesion annotations or pathological features defined by experts. Djoumessi et al. [18] compared the localization of the lesions to the annotation of the ophthalmologist and Mahipal and Kumar [21] quantitatively analyzed the CAM output using lesion annotated retinal images. These studies demonstrate the importance of lesion level overlap, Dice similarity, localization precision, fidelity, stability, and clinician agreement in addition to qualitative heatmap inspection.

The reporting of calibration and uncertainty quantification was also not consistent. High discrimination does not mean that predicted probabilities are necessarily a good reflection of clinical risk. Inaccurate confidence can lead to either over-assuring or under-referring or not escalating the cases that are not certain. Routine reporting of expected calibration error, calibration plots, Brier score, and uncertainty-aware referral strategies should therefore be considered. The evidence for fairness was even more sparse, and there was not enough subgroup data to undertake quantitative synthesis. Subsequent evaluations should include sensitivity, specificity, false-negative rates and calibration in demographic groups, devices, sites, and disease-severity strata.

4.4 Privacy, Human-AI Interaction, and Clinical Deployment

Federated learning and other distributed-learning approaches could enable distributed collaboration between institutions without the central pooling of raw retinal data [31]. However, privacy-preserving model development should not be equated with deployment readiness, as risks like gradient leakage, membership inference and institutional heterogeneity are still applicable. The clinical maturity should rather be achieved by going from algorithm development to internal validation, external validation, reader evaluation, prospective workflow validation, real-world implementation, and post-deployment monitoring. These stages are not interchangeable. External validation shows transportability, while retrospective reader studies show human-AI interaction but do not show prospective clinical benefit. Integration with referral pathway, escalation mechanism, low confidence handling, usability and monitoring for distributional change are required for clinical deployment. Performance after implementation in Indian healthcare settings is important as evidenced in post implementation studies [32] and longitudinal prediction models can be used for risk based screening intervals [33].

4.5 Clinical Governance and Regulatory Implications

Beyond accuracy, trustworthy DR-AI needs governance. The experience of regulators with autonomous retinal AI highlights the need for clear indications, validated acquisition conditions, safety mechanisms, and target populations. WHO guidance [35] also highlights autonomy, safety, transparency, accountability, inclusiveness, equity and ongoing evaluation. The EU AI [36] Act also strengthens lifecycle risk management, documentation, oversight and conformity requirements for high-risk medical AI. In this regard, regulatory readiness for DR-AI requires validated performance, traceable decision support, human oversight, monitoring of model drift, and clear accountability for decisions made with the help of AI or autonomously.

4.6 Strengths, Limitations, and Future Directions

One of the strengths of this review is the incorporation of diagnostic meta-analysis with explainability and other trustworthy-AI evaluation [34]. However, only 10 studies met the eligibility criteria and six of these studies were included in the meta-analysis. Endpoints, reference standards, and validation procedures were still heterogeneous and three 2 × 2 datasets were reconstructed from multiclass confusion matrices. Independence was further decreased by repeated use of the benchmark datasets. No prospective registration of the protocol was done and Scopus and Embase were not searched. Demographic reporting, external validation, quantitative XAI assessment, calibration, uncertainty and fairness evidence were also limited. The limited quantitative evidence base limited the ability to perform subgroup analysis and to power Deeks' asymmetry testing.

Future studies should focus on multi-center and multi-device external validation, standardization of endpoints, strict patient-level partitioning, validated explanations, calibrated uncertainty, explicit subgroup fairness analysis, privacy preserving multi-center learning, human-AI workflow studies, and continuous monitoring post-deployment. Finally, trustworthy DR-AI should be assessed as a clinical decision-support or autonomous diagnostic system in the context of a healthcare pathway, rather than as a high-performing image classifier.

5. Conclusion

Recent deep learning systems demonstrate substantial potential for Diabetic Retinopathy (DR) screening particularly for binary disease detection. The bivariate random-effects meta-analysis of the six quantitatively comparable studies showed a pooled sensitivity of 0.9842 and pooled specificity of 0.9760, corresponding to high overall diagnostic discrimination. The wide prediction region, few quantitatively comparable studies, and inconsistent study performance, however, suggest some uncertainty in transferability to future clinical settings. Multiclass severity grading was not as consistent as binary screening, and the problems of class imbalance and discrimination between adjacent disease stages persisted. Explainability was broadly integrated using CAM-based approaches, SHAP, attention, and evidence maps that are inherently interpretable, while quantitative lesion-level or clinician-based validation was still rare. Limited evidence of calibration, quantification of uncertainty, subgroup fairness, privacy-preserving learning, and prospective human-AI evaluation was also found. In addition, incomplete reporting was identified in patient-selection and reference-standard procedures in the risk-of-bias assessment. Clinical translation thus demands prospective multicentre and multi-device validation, standardised diagnostic endpoints, quantitatively and clinically validated explanations, calibrated uncertainty estimates, explicit assessment of fairness across sub-populations, privacy protection, suitable human oversight, and post-deployment monitoring. Going forward, DR-AI systems should be assessed not merely as retinal-image classifiers, but as components of clinically accountable decision-support or autonomous diagnostic pathways that rely on external validity, reliable uncertainty estimation, clinically meaningful interpretability, equitable performance, privacy protection, and accountable implementation.

Declarations

Funding Declaration

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

Competing Interests

The authors have no competing financial or non-financial interests that could have affected the work reported in this manuscript.

Ethics Statement

No new human participant recruitment, identifiable patient data collection or animal testing was conducted for this study, which was based entirely on published and publicly reported data. Thus, institutional ethics approval and informed consent were not necessary for this systematic review and meta-analysis.

Author Contribution

Sunitha T contributed to conceptualization, methodology, literature search, study screening and selection, data extraction and curation, endpoint harmonization, formal analysis, diagnostic meta-analysis, visualization, interpretation of the results, and preparation of the original manuscript draft. Ch Ram Mohan Reddy contributed to conceptualization, methodology, independent study screening and selection, verification of extracted and reconstructed data, validation of the analytical methodology and results, interpretation of the findings, supervision, project administration, and critical review and editing of the manuscript. Both authors reviewed and revised the manuscript and approved the final version for submission.

Data Availability

This study did not involve the creation of any new primary data. Data analysed in the systematic review and diagnostic meta-analysis were obtained from published studies included in the review. Study-level characteristics, diagnostic 2 × 2 data, reconstruction procedures and sensitivity analyses are presented in the main manuscript and supplementary materials.

Code Availability

The analytical code for data processing, diagnostic meta-analysis, SROC estimation, sensitivity analysis and graphical visualization is available from the corresponding author on reasonable request.

References

1. Akhtar, S. et al. A deep learning based model for diabetic retinopathy grading, *Scientific Reports*, vol. 15, Art. no. 3763, (2025). https://doi.org/10.1038/s41598-025-87171-9
2. Nunez do, J. M. et al. Rio., Using deep learning to detect diabetic retinopathy on handheld non-mydriatic retinal images acquired by field workers in community settings, *Scientific Reports*, vol. 13, Art. no. 1392, (2023). https://doi.org/10.1038/s41598-023-28347-z
3. Hassan, D. et al. Combining transfer learning with retinal lesion features for accurate detection of diabetic retinopathy, *Frontiers in Medicine*, vol. 9, Art. no. 1050436, (2022). https://doi.org/10.3389/fmed.2022.1050436
4. Zhang, W. F. et al. The validation of deep learning-based grading model for diabetic retinopathy, *Frontiers in Medicine*, vol. 9, Art. no. 839088, (2022). https://doi.org/10.3389/fmed.2022.839088
5. Arora, L. et al. Ensemble deep learning and EfficientNet for accurate diagnosis of diabetic retinopathy, *Scientific Reports*, vol. 14, Art. no. 30554, (2024). https://doi.org/10.1038/s41598-024-81132-4
6. Ikram, A. & Imran, A. ResViT FusionNet Model: An explainable AI-driven approach for automated grading of diabetic retinopathy in retinal images, *Computers in Biology and Medicine*, vol. 186, Art. no. 109656, (2025).

https://doi.org/10.1016/j.compbiomed.2025.109656

7. Dai, L. et al. A deep learning system for predicting time to progression of diabetic retinopathy. *Nat. Med.* **30** (2), 584–594. https://doi.org/10.1038/s41591-023-02702-z (2024).
8. Sharma, N. & Lalwani, P. A multi model deep net with an explainable AI based framework for diabetic retinopathy segmentation and classification, *Scientific Reports*, vol. 15, Art. no. 8777, (2025). https://doi.org/10.1038/s41598-025-93376-9
9. Gill, H. M. et al. BigEye: A clinically interpretable deep learning framework for diabetic retinopathy detection and stage prediction, *Scientific Reports*, vol. 16, Art. no. 12574, (2026). https://doi.org/10.1038/s41598-026-43573-x
10. Son, J. et al. An interpretable and interactive deep learning algorithm for a clinically applicable retinal fundus diagnosis system by modelling finding-disease relationship, *Scientific Reports*, vol. 13, Art. no. 5934, (2023). https://doi.org/10.1038/s41598-023-32518-3
11. Lepetit-Aimon, G. et al. MAPLES-DR: MESSIDOR anatomical and pathological labels for explainable screening of diabetic retinopathy, *Scientific Data*, vol. 11, Art. no. 914, (2024). https://doi.org/10.1038/s41597-024-03739-6
12. Akram, M. et al. Uncertainty-aware diabetic retinopathy detection using deep learning enhanced by Bayesian approaches, *Scientific Reports*, vol. 15, Art. no. 1342, (2025). https://doi.org/10.1038/s41598-024-84478-x
13. Abràmoff, M. D. et al. Mitigation of AI adoption bias through an improved autonomous AI system for diabetic retinal disease, *npj Digital Medicine*, vol. 7, Art. no. 369, (2024). https://doi.org/10.1038/s41746-024-01389-x
14. Wolf, R. M. et al. Autonomous artificial intelligence increases screening and follow-up for diabetic retinopathy in youth: The ACCESS randomized control trial, *Nature Communications*, vol. 15, Art. no. 421, (2024). https://doi.org/10.1038/s41467-023-44676-z
15. Joseph, S. et al. Diagnostic accuracy of artificial intelligence-based automated diabetic retinopathy screening in real-world settings: A systematic review and meta-analysis. *Am. J. Ophthalmol.* **263**, 214–230. https://doi.org/10.1016/j.ajo.2024.02.012 (2024).
16. Page, M. J. et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews, *BMJ*, vol. 372, Art. no. n71, (2021). https://doi.org/10.1136/bmj.n71
17. Rahaman, M. et al. Explainable AI for diabetic retinopathy detection using vision transformers. *Sci. Rep.* https://doi.org/10.1038/s41598-026-56569-4 (2026).
18. Djoumessi, K. et al. An inherently interpretable AI model improves screening speed and accuracy for early diabetic retinopathy, *PLOS Digital Health*, vol. 4, no. 5, Art. no. e0000831, (2025). https://doi.org/10.1371/journal.pdig.0000831
19. Ashraf, M. H. et al. HIRD-Net: An explainable CNN-based framework with attention mechanism for diabetic retinopathy diagnosis using CLAHE-D-DoG enhanced fundus images, *Life*, vol. 15, no. 9, Art. no. 1411, (2025). https://doi.org/10.3390/life15091411
20. Mutawa, A. M., Al Sabti, K., Raizada, S. & Sruthi, S. Explainable AI for diabetic retinopathy: Utilizing YOLO model on a novel dataset, *AI*, vol. 6, no. 12, Art. no. 301, (2025). https://doi.org/10.3390/ai6120301
21. Mahipal, A. & Kumar, S. Integrating explainable AI for diabetic retinopathy screening: A comparative and clinically grounded study, *Informatics in Medicine Unlocked*, vol. 63, Art. no. 101753, (2026). https://doi.org/10.1016/j.imu.2026.101753
22. Ghosh, M., Bora, V., Sahu, U. K. & Das, D. Tri-attention-augmented residual convolutional neural network with explainable AI for automated early diabetic retinopathy screening in resource-constrained environments. *Sci. Rep.* https://doi.org/10.1038/s41598-026-58716-3 (2026).
23. Civit-Masot, J. et al. An explainable ensemble for diabetic retinopathy grading with a novel confidence quality factor and configurable heatmaps. *Med. Biol. Eng. Comput.* **64** (4), 1231–1247. https://doi.org/10.1007/s11517-026-03514-2 (2026).
24. Herrero-Tudela, M. et al. An explainable deep-learning model reveals clinical clues in diabetic retinopathy through SHAP, *Biomedical Signal Processing and Control*, vol. 102, Art. no. 107328, (2025). https://doi.org/10.1016/j.bspc.2024.107328

25. Ghadai, K. K., Nayak, S. K. & Senapati, B. R. A clinically interpretable deep learning pipeline for diabetic retinopathy classification using EfficientNet, advanced data augmentation and GradCAM, *BMC Ophthalmology*, vol. 26, Art. no. 406, (2026). https://doi.org/10.1186/s12886-026-04897-4
26. Ciaramella, G. et al. A privacy-preserving and explainable method for diabetic retinopathy classification and localization. *IEEE Access.* **14**, 49830–49850. https://doi.org/10.1109/ACCESS.2026.3678201 (2026).
27. Sushith, M., Malligeswari, N., Anlin Sahaya Infant, M., Tinu & Jaiganesh, M. HyperGraph-based capsule temporal memory network for efficient and explainable diabetic retinopathy detection in retinal imaging, *Scientific Reports*, vol. 16, Art. no. 607, (2026). https://doi.org/10.1038/s41598-025-30128-9
28. Sushith, M., Sathiya, A., Kalaipoonguzhali, V. & Sathya, V. A hybrid deep learning framework for early detection of diabetic retinopathy using retinal fundus images, *Scientific Reports*, vol. 15, Art. no. 15166, (2025). https://doi.org/10.1038/s41598-025-99309-w
29. Wang, Z., Wang, Y., Ma, C., Bao, X. & Li, Y. Diabetic retinopathy classification using a multi-attention residual refinement architecture, *Scientific Reports*, vol. 15, Art. no. 29266, (2025). https://doi.org/10.1038/s41598-025-15269-1
30. Abbasi, R. et al. Diabetic retinopathy detection using adaptive deep convolutional neural networks on fundus images, *Scientific Reports*, vol. 15, Art. no. 24647, (2025). https://doi.org/10.1038/s41598-025-09394-0
31. Tashrif, M. T. A. et al. Privacy-aware diabetic retinopathy grading and visual lesion-focused interpretability through mixture-of-experts federated deep learning with explainable AI. *Sci. Rep.* https://doi.org/10.1038/s41598-026-58292-6 (2026).
32. Brant, A. et al. Performance of a deep learning diabetic retinopathy algorithm in India, *JAMA Network Open*, vol. 8, no. 3, Art. no. e250984, (2025). https://doi.org/10.1001/jamanetworkopen.2025.0984
33. Nderitu, P. et al. Predicting 1, 2 and 3 year emergent referable diabetic retinopathy and maculopathy using deep learning, *Communications Medicine*, vol. 4, Art. no. 167, (2024). https://doi.org/10.1038/s43856-024-00590-z
34. Naz, H., Ahuja, N. J. & Nijhawan, R. Diabetic retinopathy detection using supervised and unsupervised deep learning: A review study, *Artificial Intelligence Review*, vol. 57, Art. no. 131, (2024). https://doi.org/10.1007/s10462-024-10770-x
35. World Health Organization. Ethics and Governance of Artificial Intelligence for Health: WHO Guidance (World Health Organization, 2021); https://www.who.int/publications/i/item/9789240029200
36. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) (Official Journal of the European Union, 2024); https://eur-lex.europa.eu/eli/reg/2024/1689/oj
37. Whiting, P. F. et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann. Intern. Med.* **155**, 529–536 (2011).
38. McInnes, M. D. F. et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. *JAMA* **319**, 388–396 (2018).
39. Deeks, J. J., 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.* **58**, 882–893 (2005).
40. Rutter, C. M. & Gatsonis, C. A. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat. Med.* **20**, 2865–2884. https://doi.org/10.1002/sim.942 (2001).
41. Reitsma, J. B. et al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J. Clin. Epidemiol.* **58**, 982–990. https://doi.org/10.1016/j.jclinepi.2005.02.022 (2005).

Figures

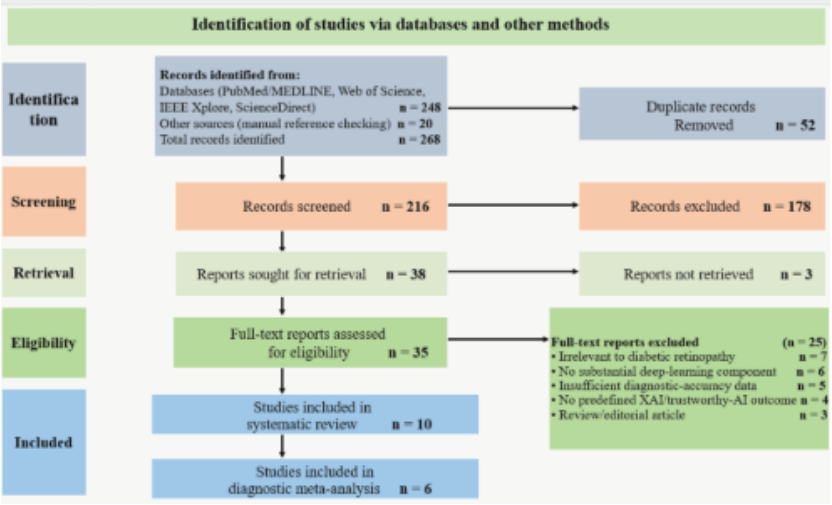

Figure 1

PRISMA 2020 flow diagram illustrating study identification, screening, eligibility assessment, and inclusion. A total of 268 records were identified from four electronic databases and manual reference screening. After removing 52 duplicate records, 216 unique records were screened, 35 studies underwent full-text eligibility assessment, 10 studies were included in the systematic review, and 6 studies met the eligibility criteria for the diagnostic meta-analysis.

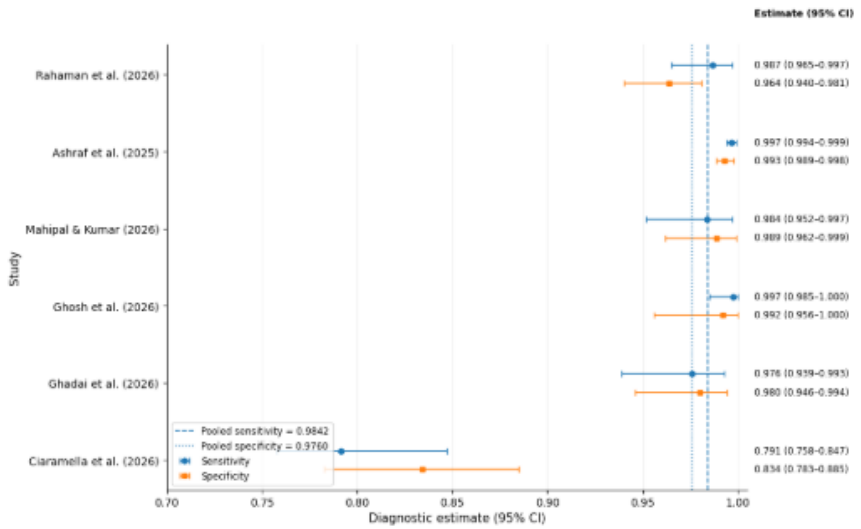

Figure 2

Forest plot of study-specific sensitivity and specificity with 95% confidence intervals for the six studies included in the diagnostic meta-analysis.

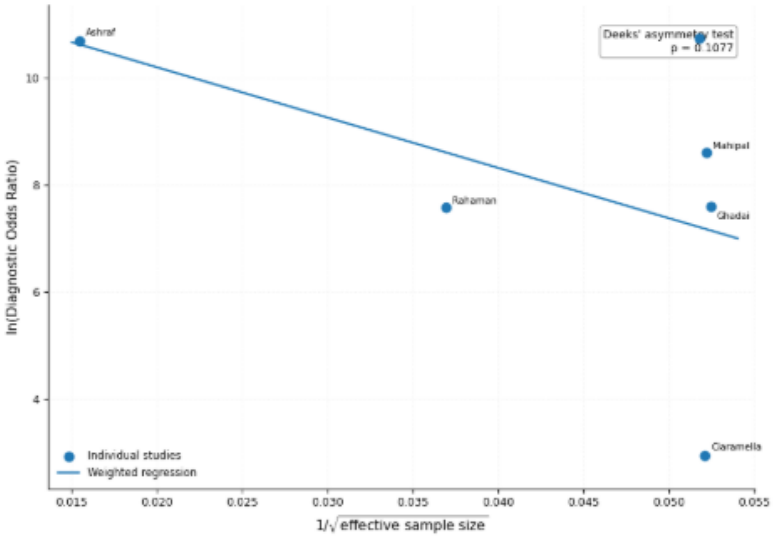

Figure 3

Exploratory Deeks' funnel plot for small-study effects (p = 0.1077).

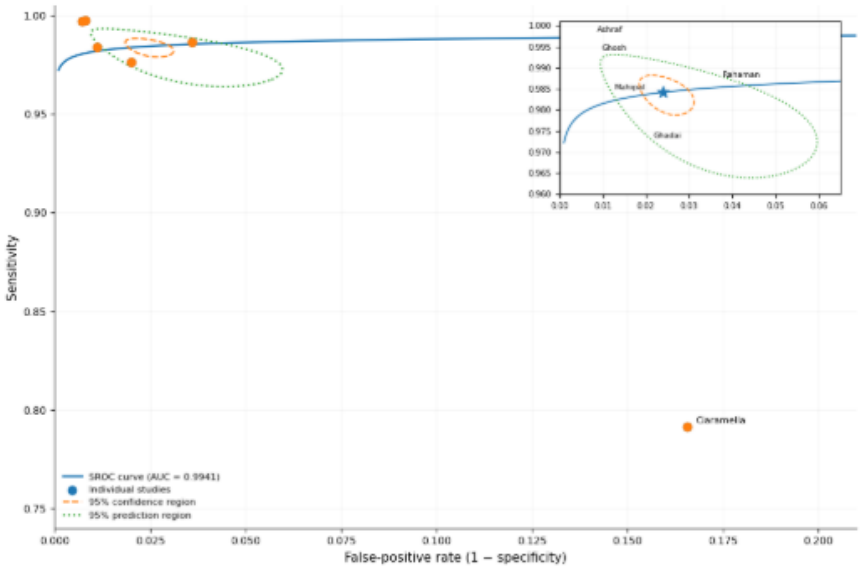

Figure 4

Summary receiver-operating-characteristic curve showing study-level operating points, pooled summary estimate, 95% confidence region, and 95% prediction region.

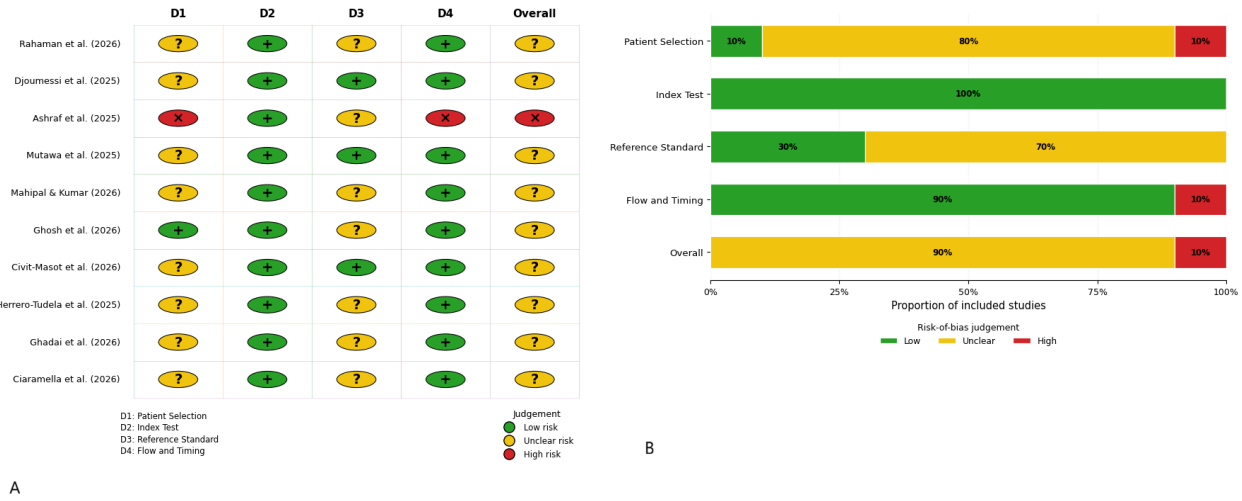

Figure 5

A. QUADAS-2 traffic-light assessment of study-level risk of bias.

B. Distribution of QUADAS-2 judgments across risk-of-bias domains.

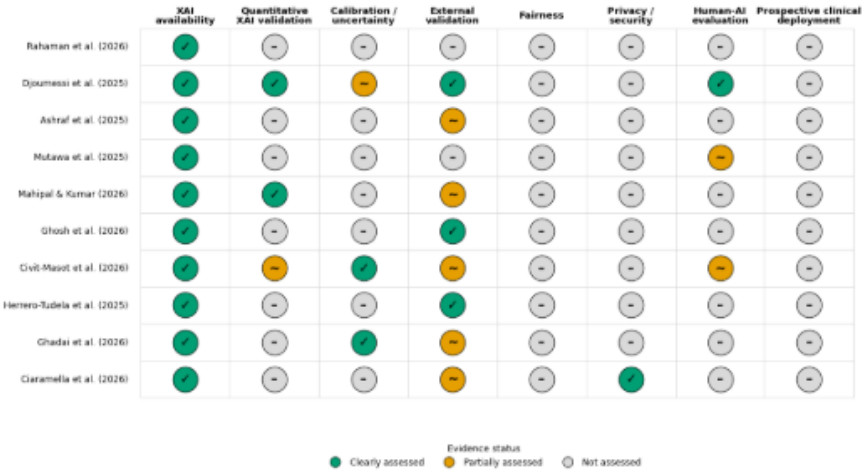

Figure 6

Integrated evidence map of explainability and trustworthy-AI characteristics across the included DR studies.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTables.docx