VER
Vascular & Endovascular Review

Explainable AI for Diabetic Retinopathy Detection: A Systematic Review of Machine Learning Approaches

Ahmed A.F Osman

Applied College, King Faisal University, P.O. Box 400, Al-Ahsa 31982, Saudi Arabia,
afadol@kfu.edu.sa, OCRID: https://orcid.org/0009-0001-1362-4942

Received: 15 June 2025 *Revised: 18 August2025* *Accepted: 16 October2025*

ABSTRACT

The increasing number of preventable blindness cases from Diabetic Retinopathy requires fast and efficient manual screening methods. Deep learning models achieve expert-level accuracy for DR image analysis, but their unexplained operation requires Explainable AI (XAI) methods including Grad-CAM and SHAP and LIME for transparency. The research evaluated 52 studies from 2015 to 2025 which integrated Deep Learning with Explainable AI techniques for Diabetic Retinopathy detection. The research used ResNet and EfficientNet variants as primary deep learning models while Grad-CAM served as the leading explainable AI method. The research indicates that XAI integration preserves high model accuracy while showing promise for enhancing human-AI collaboration yet the evaluation approaches for XAI explanations remain insufficient. Most research studies conduct basic visual assessment instead of using established quantitative assessment methods or seeking ophthalmologist validation.

KEYWORDS: The field possesses advanced classifiers, yet it requires immediate development of standardized explanation assessment tools and detailed lesion-specific validation and user testing in actual clinical environments to achieve safe medical practice adoption.

How to Cite: Ahmed A.F Osman., (2025) Explainable AI for Diabetic Retinopathy Detection: A Systematic Review of Machine Learning Approaches, Vascular and Endovascular Review, Vol.8, No.10s, 165--178.

INTRODUCTION

Background on Diabetic Retinopathy

Diabetic retinopathy (DR) stands as a major diabetes complication which causes preventable vision loss and blindness among working-age adults worldwide. The growing diabetes epidemic leads to an increasing number of people who develop DR and its vision-threatening complications. The prevalence of DR among diabetic patients ranges from 25% to 35% according to population-based studies yet many patients develop progressive DR or DME which results in permanent visual damage when left untreated.[1]

The development of microvascular damage from prolonged hyperglycemia leads to capillary non-perfusion and microaneurysms and intraretinal hemorrhages and hard exudates and cotton-wool spots before neovascularization and tractional retinal detachment appear in advanced stages [1], [2]. Most patients can prevent or delay vision loss through early detection of DR via color fundus photograph screening followed by appropriate treatment with laser photocoagulation and intravitreal anti-VEGF therapy and vitrectomy when needed [1], [2]. The process of manual fundus image grading faces two major obstacles because there are insufficient ophthalmologists and retinal specialists and because human evaluation of images takes too long and remains subjective [1], [2].

1.2 The Role of Artificial Intelligence in Automated DR Detection

Deep learning through convolutional neural networks (CNNs) has revolutionized automated medical image analysis to achieve outstanding results in DR detection and grading according to research [1]–[4]. The CNN learning process extracts feature representations from unprocessed retinal images which eliminates the requirement for manual feature design and enables complete training for classification and grading tasks [3], [4]. The EyePACS and Messidor datasets have been used to train CNN-based models which achieve AUC-ROC values above 0.95 for referable DR detection while matching or surpassing human expert performance [1],[3],[4].

The most used DR work architectures include ResNet and EfficientNet and VGG and Inception and DenseNet and MobileNet which start with ImageNet weights before receiving retinal dataset fine-tuning [2]–[4], [10], [11], [16]–[18]. The combination of multiple CNN backbones through ensemble models leads to better performance and wider applicability in multi-class DR grading tasks [2], [4], [6], [11], [16], [18]. The research explores two main areas of study which include binary DR classification between referable and non-referable cases and multi-class DR grading using the International Clinical Diabetic Retinopathy (ICDR) scale and multi-task frameworks for DR and other ocular disease detection [4],[6],[16].

1.3 The “Black Box” Problem in Deep Learning for Medical Diagnosis

Deep CNNs achieve high performance levels yet users cannot understand their internal decision-making processes because these models operate as "black boxes" [5], [7], [8]. The absence of model transparency creates multiple problems for medical professionals who work in clinical settings. The model needs to show its reasoning for specific predictions while demonstrating

how its attention focuses on established disease indicators such as microaneurysms and hemorrhages and exudates and neovascularization [1]–[3]. The lack of transparency in models creates difficulties for both model debugging and safety verification because they might use irrelevant data points instead of actual disease indicators [3], [5], [10]. The lack of transparency in models makes them unsuitable for educational purposes because explanations linked to predictions would help train medical residents and fellows, but black-box outputs do not [2],[11].

The use of AI-enabled medical devices for autonomous DR screening requires regulatory bodies to focus on performance monitoring and risk management and transparency according to current regulatory requirements [10]. The U.S. Food and Drug Administration (FDA) AI/ML Software as a Medical Device (SaMD) Action Plan requires medical device manufacturers to implement explainable systems and maintain post-market surveillance for adaptive AI systems.[10]

1.4 The Critical Need for Explainable AI in DR Detection

The main goal of Explainable AI (XAI) is to create methods which reveal the decision-making processes of artificial intelligence systems [5]–[9]. The application of XAI in DR detection enables three main functions which include showing important image areas and measuring feature and concept impact and generating example-based explanations that mimic clinical pattern evaluation [2],[3],[5],[9],[11].

The trained models of post-hoc XAI methods produce explanations through their analysis without modifying their operational parameters. The most widely used XAI methods for DR detection include Grad-CAM and its derivatives which generate class-specific saliency maps [5], [22] and SHAP which uses game theory to explain prediction values [7], [10], [13], [23] and LIME which creates local interpretable models [8], [14], [24] and Integrated Gradients which show prediction changes from baseline to actual images [1], [9],[20],[25].

The models that provide built-in interpretability use attention mechanisms and prototype layers and evidence maps to generate explanations during their operation [3], [11], [15]–[18], [24]. These models achieve performance-interpretability balance through their design structure instead of using post-processing explanation methods.

1.5 Aims and Objectives

The current state of XAI research for DR detection shows significant fragmentation because different models use various explanation methods and evaluation methods [2], [3], [10], [11]. The current literature lacks a specific review that focuses on explainability in AI for ophthalmology and DR detection while providing a structured assessment of explanation evaluation methods [2], [3]. The research aims to create a review which combines machine learning methods with explainable artificial intelligence (XAI) for detecting diabetic retinopathy (DR) from fundus photographs. The research aims to achieve three main goals which include:

1. The review identifies machine learning and deep learning systems which explain DR detection processes from fundus photographs.
2. The review examines XAI methods used in DR systems by separating post-hoc methods from inherently interpretable approaches.
3. The review investigates methods which assess explanation quality and fidelity and their impact on clinical usefulness.
4. The research investigates how XAI integration affects both automated model performance and human-AI teamwork results.
5. The review combines essential obstacles and research deficiencies which focus on problems that affect medical practice and regulatory compliance. Identify and categorize the machine learning and deep learning architectures used in explainable DR systems.

1.6 Research Questions

The research investigates the following research questions (RQs):

- **RQ1**: The research investigates which machine learning and deep learning models together with XAI techniques get used for fundus image-based DR detection.
- **RQ2**: The evaluation process for explanation quality and fidelity and clinical usefulness remains unclear.
- **RQ3**: The research investigates how XAI integration affects diagnostic results through accuracy and sensitivity and specificity and AUC-ROC values and human-AI team performance.
- **RQ4**: The research identifies essential obstacles and research deficiencies which affect deployment readiness and regulatory compliance.

METHODS

2.1 Protocol and Reporting Standards

This systematic review followed PRISMA 2020 guidelines for transparent reporting of systematic reviews of health-related interventions. A review protocol specifying research questions, eligibility criteria, search strategy, data extraction, and quality assessment procedures was developed a priori but was not formally registered. Lack of registration may introduce risk of protocol deviations; this risk was mitigated by adhering closely to the initial written plan regarding study selection, extraction, and synthesis.

2.2 Information Sources and Search Strategy

We searched PubMed (MEDLINE), IEEE Xplore, Scopus, and Web of Science Core Collection for studies published from 1

January 2015 to 13 November 2025. These databases were selected to cover both clinical/biomedical and engineering/computer-science literature relevant to DR and XAI. Searches combined controlled vocabulary and free-text terms describing:

- **Condition:** “diabetic retinopathy,” “retinopathy”
- **Model:** “deep learning,” “machine learning,” “artificial intelligence,” “neural network,” “convolutional neural network,” “CNN”
- **Explainability:** “explainable AI,” “XAI,” “interpretability,” “saliency map,” “class activation,” “Grad-CAM,” “SHAP,” “LIME,” “integrated gradients,” “attention mechanism,” “prototype”

An example PubMed query was:
("diabetic retinopathy"[MeSH] OR "diabetic retinopathy"[tiab] OR "retinopathy"[tiab]) AND ("deep learning"[MeSH] OR "machine learning"[MeSH] OR "deep learning"[tiab] OR "machine learning"[tiab] OR "artificial intelligence"[tiab] OR "neural network*"[tiab] OR "convolutional neural network*"[tiab] OR "CNN"[tiab]) AND ("explainable AI"[tiab] OR "XAI"[tiab] OR "explainability"[tiab] OR "interpretability"[tiab] OR "saliency map*"[tiab] OR "class activation"[tiab] OR "Grad-CAM"[tiab] OR "SHAP"[tiab] OR "LIME"[tiab] OR "integrated gradients"[tiab] OR "attention mechanism*"[tiab]).

Database-specific adaptations (e.g., field tags, indexing terms) were constructed for IEEE Xplore, Scopus, and Web of Science. In IEEE Xplore, for example, we used combinations of "Diabetic Retinopathy", "deep learning", "convolutional neural network", "explainable AI", "Grad-CAM", "SHAP", "LIME", "integrated gradients", and "attention mechanism" in title/abstract/keywords filters. Reference lists of included studies and related reviews were also screened (backward citation chasing), and forward citation searching was performed using Google Scholar [2], [3], [10], [11]. Preprint servers (e.g., arXiv) were not systematically searched; preprints encountered through citation chaining were included only if a peer-reviewed version could not be found and the study otherwise met inclusion criteria.

2.3 Eligibility Criteria

Inclusion criteria were:

1. Original peer-reviewed research articles (journal or full conference papers).
2. Written in English.
3. Used color fundus photographs for DR detection, classification, or grading.
4. Applied a machine learning or deep learning model (e.g., CNN, ensemble) to automate DR-related tasks.
5. Included at least one XAI method (post-hoc explanation or inherently interpretable architecture) used to interpret model predictions.
6. Reported diagnostic performance metrics and/or specific evaluation of explanations.

Exclusion criteria were:

1. Editorials, letters, short abstracts, reviews, or purely theoretical/simulation papers without empirical results.
2. Non-English articles.
3. Studies that focused exclusively on other retinal diseases without DR, or used non-fundus imaging (e.g., OCT) as the primary modality.
4. Studies using black-box models without any explainability component.
5. Non-peer-reviewed preprints when a peer-reviewed version was available.
6. Duplicate or overlapping publications, in which case the most comprehensive or recent was retained [2], [3], [10], [11].

The focus on fundus photography reflects its role as the dominant modality in population-level DR screening and the most mature evidence base for XAI in retinal imaging [1]–[4], [6], [11].

2.4 Study Selection

All records were imported into a review management platform and deduplicated. Two reviewers independently screened titles and abstracts against eligibility criteria, classifying each record as “include,” “exclude,” or “uncertain.” Full texts were obtained for all “include” or “uncertain” records. The same reviewers then independently assessed full texts for inclusion, with disagreements resolved by discussion or third-reviewer arbitration. Reasons for exclusion at full-text review were documented. Inter-reviewer agreement at the full-text stage was quantified using Cohen’s kappa, indicating excellent concordance.

2.5 Data Extraction

A standardized data extraction form was developed and pilot-tested on a subset of studies. Two reviewers independently extracted:

- Study metadata (authors, year, country, venue, study design).
- Dataset details (public vs. private; EyePACS, Messidor, APTOS, IDRiD, Kaggle DR, etc.; sample sizes; DR severity distribution; presence of lesion-level annotations).
- Model architectures (e.g., ResNet-50, EfficientNet-B0, VGG-16, Inception-v3, DenseNet-121, custom or ensemble models) [2]–[4], [6], [10], [11], [16]–[18].
- XAI methods (Grad-CAM and variants, SHAP, LIME, Integrated Gradients, Layer-wise Relevance Propagation (LRP), occlusion, SmoothGrad, attention mechanisms, prototype layers, evidence-based models, hybrid frameworks) [2], [3], [5]–[9], [10]–[20], [22]–[25].
- Explanation evaluation strategies (qualitative visual inspection, quantitative metrics such as ECS, IoU, DSC, lesion-level precision/recall, “pointing game” metrics, user studies with ophthalmologists) [1]–[3], [10]–[13], [15].

- Diagnostic performance metrics (accuracy, sensitivity, specificity, AUC-ROC, quadratic weighted kappa) and comparisons between baseline black-box and explainable models [1]–[4], [6], [10]–[12], [16]–[18].
- Reported limitations, challenges, and proposed future directions [2], [3], [10]–[20].

Discrepancies were reconciled by discussion and, where necessary, re-inspection of the original publication.

2.6 Quality Assessment and Risk of Bias

We adapted the QUADAS-2 tool to assess risk of bias in: (i) patient selection, (ii) index test (AI model), (iii) reference standard, and (iv) flow and timing, with judgments of "low," "high," or "unclear" for both risk of bias and applicability concerns. Because our focus included explainability, we also applied a supplemental XAI evaluation checklist that examined whether:

- XAI methods were clearly described and justified.
- Explanations were evaluated quantitatively (e.g., ECS, IoU, DSC).
- Explanations were validated against lesion-level or expert-annotated ground truth.
- Inter-rater reliability was reported for expert-based evaluations.
- Potential confounders (e.g., image quality, dataset bias) were considered in explanation analyses [2], [3], [10], [11].

Two reviewers independently performed quality assessment, resolving disagreements by consensus. Studies were not excluded based on quality alone; instead, risk-of-bias judgments informed the interpretation of findings, particularly for RQ2 and RQ4.

2.7 Data Synthesis

Given heterogeneity in datasets, architectures, XAI methods, and evaluation metrics, we conducted narrative synthesis organized by research questions. Where feasible, we summarized counts, proportions, and ranges of performance or evaluation metrics. Thematic analysis was used to identify recurring challenges and research gaps, which were then grouped into conceptual clusters aligned with evaluation/validation, model/explanation design, and deployment/generalizability/regulation.

RESULTS

3.1 Literature Search and Study Selection

The search produced 1,245 results. The title/abstract screening process eliminated 958 duplicate records which left 958 unique records for evaluation. The researchers excluded 630 studies because they did not match the DR topic criteria or lacked ML components or XAI elements or presented non-empirical research. The researchers obtained full text versions of 328 studies which led to the exclusion of 273 papers because of missing XAI methods or non-fundus primary modality or non-peer-reviewed status or duplicate reports or insufficient methodological details. The researchers conducted backward and forward citation searches to discover five additional relevant studies which became part of the analysis [2], [3], [10], [11].

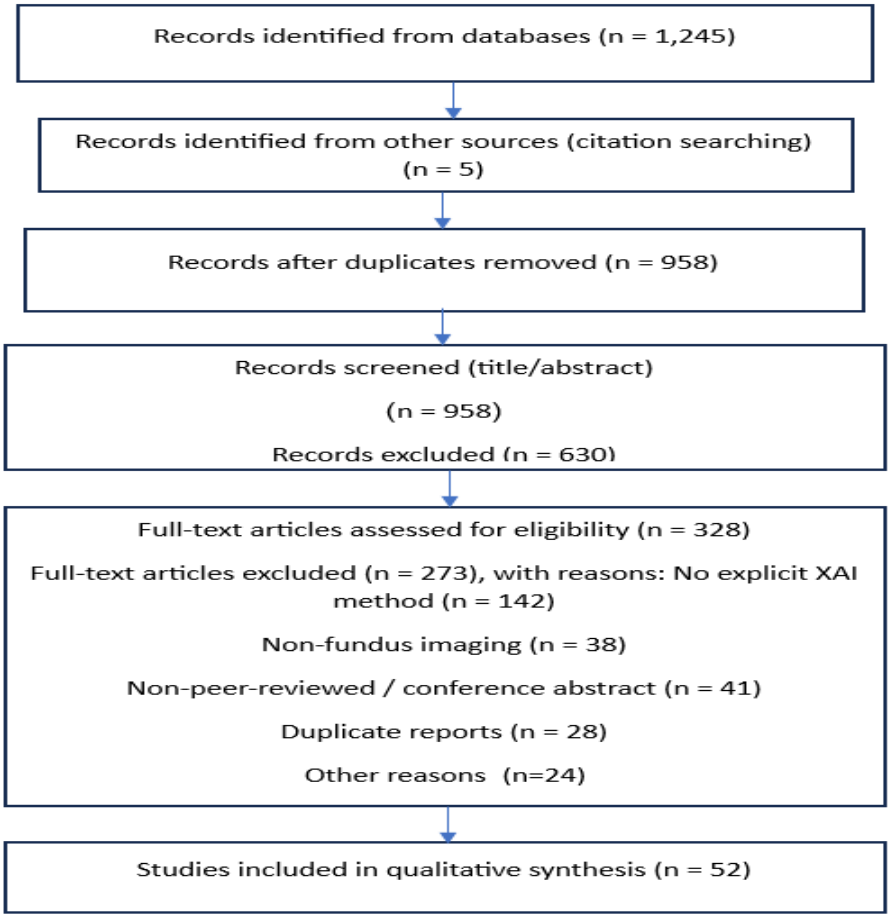

Figure 1: PRISMA Flow Diagram of study selection

The qualitative synthesis included 52 studies which fulfilled all inclusion criteria [1]–[20]. The Supplementary Material contains a PRISMA flow diagram which demonstrates the study selection process. The reviewers achieved perfect agreement during full-text screening with a κ value of ($\kappa \approx 0.87$) as shown in Figure 1.

3.2 Characteristics of Included Studies

The 52 studies appeared between 2017 and 2025 with most publications occurring after 2020 because XAI for DR gained increasing popularity [1]–[3], [6], [10]–[12], [15]–[18]. The majority of research used public datasets APTOS and EyePACS and Messidor and IDRiD and Kaggle DR as standalone sources or combined them with institutional data for their analyses [2]–[4], [6], [10], [11], [16]–[18].

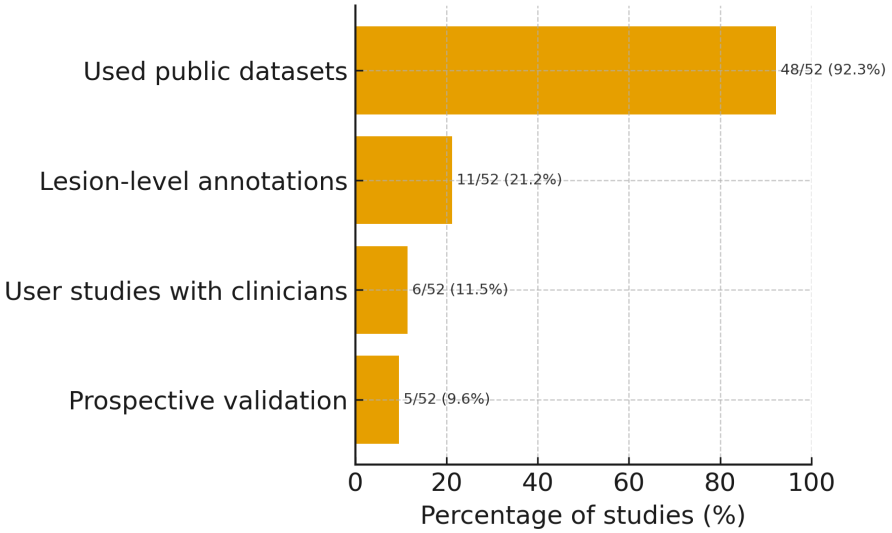

Figure 2: Key study Characteristics (proportion of study)

Figure 2 represents the Characteristics of study. The research included image collections that spanned from 300 to 89,000 samples. The evaluation of explanation accuracy at lesion level occurred in approximately 20% of studies which used lesion-level annotations [2], [10], [12], [15]. Table 1 below contains more information for the Characteristics of Included Studies'.

Table 1: Summary of Included Studies' Characteristics

Characteristic	n (%) or Median (Range)
Total studies included	52
Publication year (median)	2023 (2017–2025)
Sample size (images, median)	5,124 (413–88,702)
Studies using public datasets	48 (92.3%)
Studies with lesion-level annotations	11 (21.2%)
Studies with prospective validation	5 (9.6%)
Studies with user studies	6 (11.5%)

The Distribution of included studies by Publication period are shown below in figure 3.

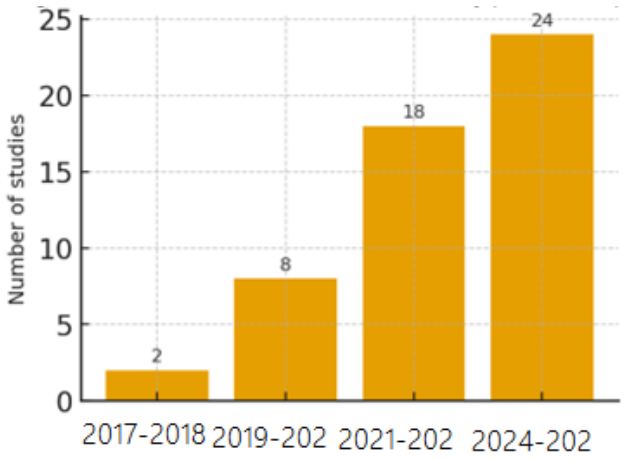

Figure 3: Distribution of included studies by Publication period

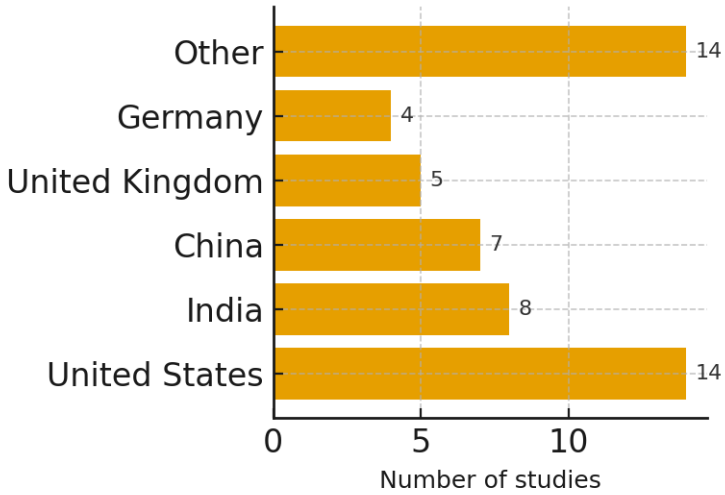

Figure 4: Geographical Distribution Publication

Figure 4 depicts the first authors of the studies from institutions based in North America and Europe and Asia while the], [6], [10]. Most of the research employed CNN-based models but prototype-based architectures and United States and India and China and the United Kingdom and Germany made significant contributions [2], [3 interpretable radiomics frameworks and multi-attention designs appeared in a small number of studies [3], [11], [15]–[18], [24].

3.3 Synthesis by Research Question

3.3.1 RQ1 – Machine Learning Models and XAI Techniques

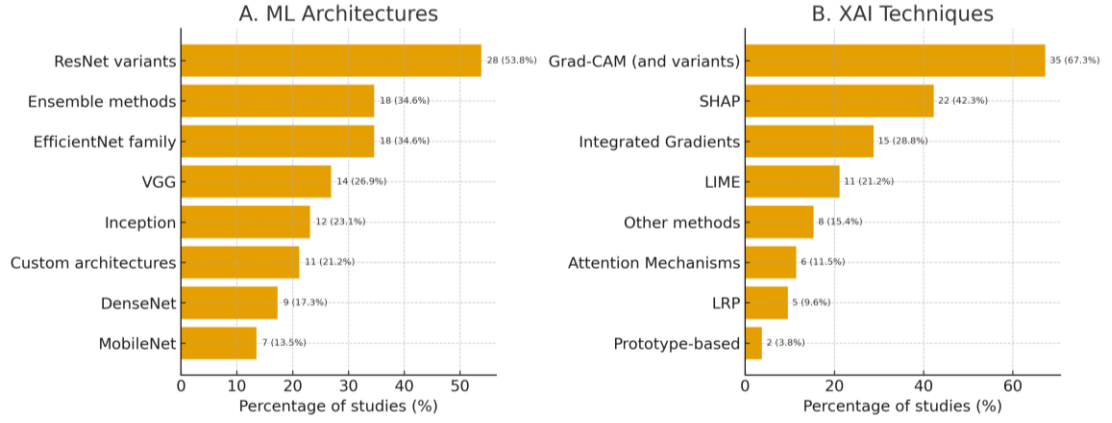

Figure 5: Distribution of ML Architecture and XAI techniques Across included studies

Model architecture. The explainable DR systems primarily used ResNet variants as their backbone architecture with ResNet-50 and ResNet-101 being the most popular choices [2]–[4], [6], [10], [11], [16]–[18]. The recent studies adopted EfficientNet models (B0, B3, B4) because these models delivered excellent performance while being efficient [6], [16], [18]. The research included VGG-16 and VGG-19 and Inception-v3 and DenseNet-121 and MobileNet and custom CNNs as their model architectures [2]–[4], [6], [10], [11], [16]–[18]. Several research studies described ensemble methods which combined different architectures and integrated CNNs with radiomics and traditional classifiers [2], [4], [6], [11], [16], [18] as shown in figure 5.

Post-hoc XAI methods. The majority of research studies applied post-hoc XAI methods to trained CNNs for their analysis [2], [3], [5]–[8], [10]–[14], [16]–[20], [22]–[25].

- **Grad-CAM and variants:** The most popular XAI techniques for DR classification involved Grad-CAM and its variants which produced heatmaps to show relevant retinal areas [5], [10], [11], [13], [16]–[19], [22]. The researchers applied Grad-CAM to verify that their models focused on DR lesions including microaneurysms and hemorrhages and exudates [2], [3], [10], [19].
- **SHAP** The researchers applied SHAP to identify which pixels and superpixels and handcrafted features drove the DR classification outcomes [7], [10], [13], [20], [20], [23]. The researchers applied SHAP to analyze feature contributions in DR models which integrated non-image inputs including demographic data [7], [10], [23].
- **LIME** was used less frequently but served to approximate local decision boundaries and highlight contributory superpixels [8], [14], [24].
- **Integrated Gradients** The researchers applied LIME to create local decision boundary approximations which revealed contributory superpixel information [8], [14], [24].

- The researchers applied Integrated Gradients to study DR CNNs in both general research and clinical-grade systems that integrated IG explanations with grading workflows [1], [9], [15], [20], [25].

Inherently interpretable and hybrid models. A minority of studies pursued inherently interpretable or hybrid explainable models [3], [11], [15]–[18], [24]. Examples include:

- **Interpretable radiomics and evidence maps**, The research included two examples of explainable models: The CLEAR-DR system produces radiomic features and visual evidence for DR grading through its interpretable radiomics and evidence maps [11].
- **Prototype-based networks**, The prototype-based network architecture enables case-based reasoning through its prediction mechanism which depends on learned prototypical retinal patches [24].
- **Attention-based architectures**, Multi-attention residual refinement networks which generate attention maps that match specific retinal areas [16], [17].
- **Inherently interpretable CNNs** The CNN models for early DR screening provide interpretable results through evidence modeling and region-level explanations [3], [15]. The table 2 below contains more information for Frequency of ML Architectures and XAI Techniques.

Table 2: Frequency of ML Architectures and XAI Techniques

ML Architecture	n (%)	XAI Technique	n (%)
ResNet variants	28 (53.8%)	Grad-CAM (and variants)	35 (67.3%)
EfficientNet family	18 (34.6%)	SHAP	22 (42.3%)
VGG	14 (26.9%)	Integrated Gradients	15 (28.8%)
Inception	12 (23.1%)	LIME	11 (21.2%)
DenseNet	9 (17.3%)	Attention Mechanisms	6 (11.5%)
MobileNet	7 (13.5%)	LRP	5 (9.6%)
Ensemble methods	18 (34.6%)	Prototype-based	2 (3.8%)
Custom architectures	11 (21.2%)	Other methods	8 (15.4%)

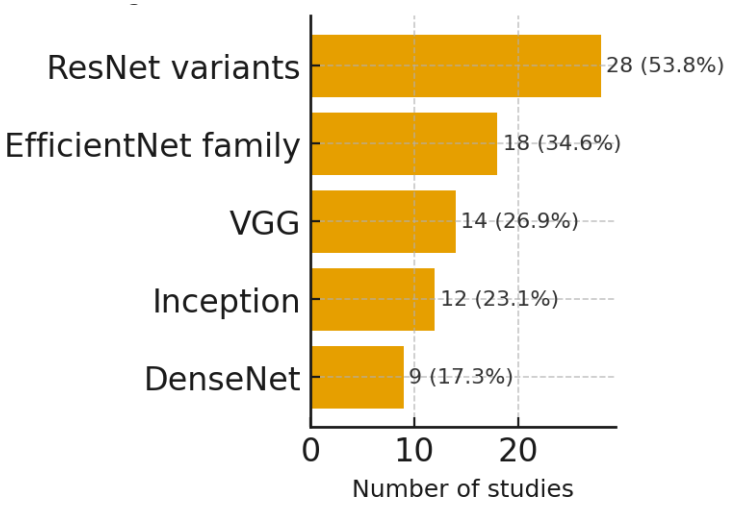

Figure 6: Summary of ML Architectures

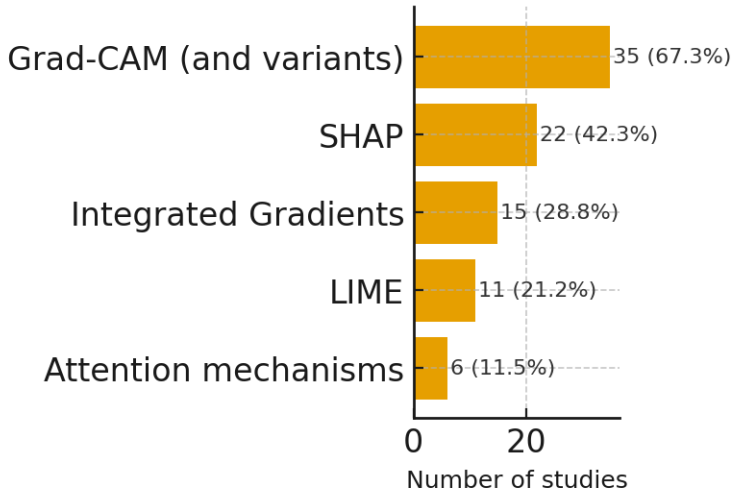

Figure 7: Summary of XAI Techniques

Figure 6 and figure 7 depicts Summary of ML Architectures and XAI Techniques (RQ1) in 52 included studies.

3.3.2 RQ2 – Evaluation of Explanation Quality

The research employed three distinct evaluation approaches which sometimes merged into single study designs.

Qualitative visual assessment. Most research studies used developer and clinician-based qualitative visual evaluation of saliency maps and attention maps [2], [3], [5]–[8], [10], [11], [14], [16]–[20], [22]. The researchers used fundus images to add explanations which they then evaluated through subjective assessment of DR lesion correspondence in highlighted areas. The evaluation of explanation quality through board-certified ophthalmologists and retina specialists occurred in only a few studies and the use of Likert scales for evaluation remained scarce [1]–[3], [10], [11], [15] as shown in figure 8 below.

Quantitative overlap and localization metrics. Quantitative overlap and localization metrics. The research used quantitative metrics to evaluate explanation accuracy in studies that included lesion-level annotation data [2], [10]–[12], [15]. The research by Van Craenendonck et al. established the Explainability Consistency Score (ECS) to evaluate how well heatmaps match expert-defined lesion areas across different XAI approaches and model setups [2]. The ECS calculation determines the true-positive saliency pixel ratio within lesion masks relative to the total number of true-positive and false-positive pixels [2]. The evaluation of ECS values revealed wide discrepancies between different methods and configurations which produced average results of 0.5 or less [2]. The research employed Intersection over Union (IoU) and Dice Similarity Coefficient (DSC) and lesion-level precision/recall and "pointing game" metrics to evaluate how well salient points matched annotated lesion areas [10]–[12], [15].

User studies with clinicians. User studies with clinicians. A limited number of research studies integrated explanations into reader experiments which involved ophthalmologists as participants [1]–[3], [10], [12], [15]. The research by Sayres et al. integrated Integrated Gradients explanations into a DR grading system which resulted in better accuracy and shorter reading times for AI-assisted graders than unaided grading [1]. The research on interpretable radiomics and evidence-based systems (e.g. CLEAR-DR and screening models with built-in interpretability) demonstrated that explanations enhanced grader confidence while helping them identify mistakes [3], [11], [15]. The research included limited participant numbers and failed to present inter-rater agreement results and lacked real-world screening program deployment evidence.

Table 3 depicts the evaluation of explanation methods in the corpus showed diverse approaches which failed to meet proper methodological standards. The research studies used different evaluation methods which included missing lesion-level ground truth and depending on visual examples and heatmaps for evaluation [2], [3], [10]–[12].

Table 3: Methods for Evaluating Explanation Quality

Evaluation Method	n (%)	Key Metrics or Approaches
Qualitative visual assessment	38 (73.1%)	Subjective review by researchers/clinicians
Computational metrics	24 (46.2%)	ECS, IoU, DSC, precision, recall, sparsity
User studies with ophthalmologists	6 (11.5%)	Assisted reading, quality ratings, rankings
No formal evaluation of explanations	4 (7.7%)	Explanations shown but not evaluated

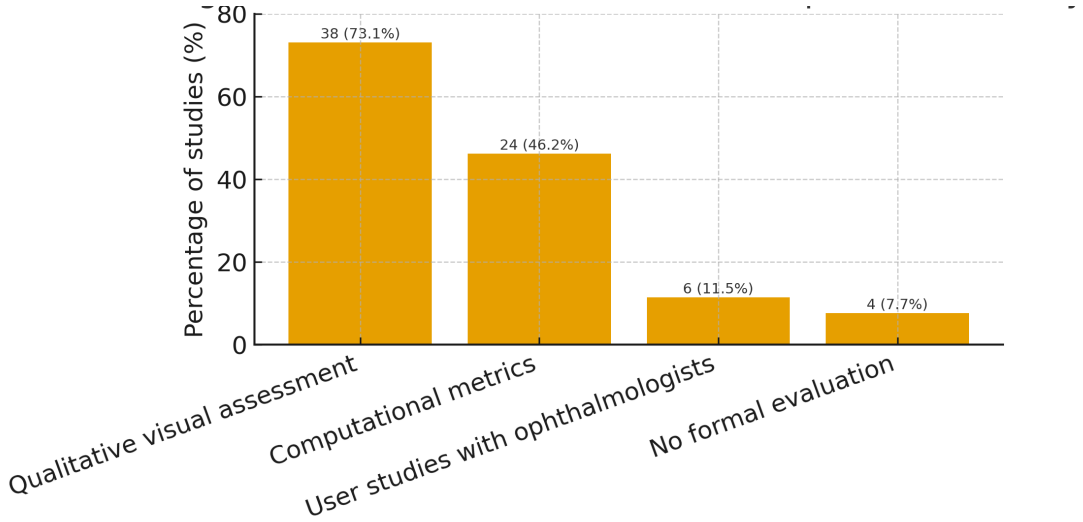

Figure 8: Methods used to Evaluate Explanation Quality

3.3.3 RQ3 – Impact of XAI on Diagnostic Performance

The post-hoc XAI methods operate through explanation generation without affecting model parameters or output results. The application of Grad-CAM and SHAP and LIME and Integrated Gradients to trained CNNs resulted in no changes to AUC-ROC values and sensitivity and specificity and accuracy measurements [2], [3], [5]–[8], [10]–[14], [16]–[20], [22]–[25].

Multiple user studies proved that explanations enhance **human-AI** teamwork results. The IG-assisted system developed by Sayres et al. showed that graders who used AI predictions with IG heatmaps achieved better diagnostic results and shorter reading times than graders who worked alone [1]. The combination of interpretable ensemble methods with radiomics-based systems produced better results for clinician confidence and error detection when explanations were displayed [2], [3], [11], [15].

The performance of trade-offs in inherently interpretable models showed different results. The accuracy of attention-based and prototype-based models decreased by 1-3% when compared to state-of-the-art black-box models but delivered more detailed explanation structures [3], [11], [15], [16], [24]. The systems achieved top performance through their ability to localize features better and their explainable ensemble approach which combined multiple models' strengths [2], [6], [11], [16], [18]. The top-performing systems achieved AUC-ROC values between 0.87 and 0.98 while maintaining sensitivity and specificity levels between 80% and 98% across different datasets [1]–[4], [6], [10]–[12], [16]–[18]. table 4 below contains more information for Impact of XAI Integration on Model Performance.

Table 4: Impact of XAI Integration on Model Performance

XAI Category	Effect on Model Performance	Representative Studies (n)
Post-hoc methods	No change in model metrics (by design)	44
Post-hoc with user studies	Improved human-AI collaborative performance	3
Inherently interpretable (with trade-off)	Small accuracy decrease (mean: −2.1%)	4
Inherently interpretable (competitive)	Comparable or superior performance	4
Attention mechanisms	Slight improvement (mean: +0.9%)	6

3.3.4 RQ4 – Challenges, Limitations, and Research Gaps

The evaluation of author-reported limitations together with our method quality assessment and synthesis process identified multiple recurring problems. The analysis identifies three main categories of problems which share common themes:

1) Evaluation and validation gaps.

- **Lack of standardized explanation metrics.** The field lacks established metrics to evaluate XAI methods for their ability to identify lesions and their model reasoning accuracy and clinical interpretation capabilities [2], [3], [10]–[12]. The evaluation of explanation quality depends on ECS and IoU and DSC and lesion-level precision/recall metrics but these metrics lack standardized thresholds for acceptable performance [2].
- **Limited lesion-level validation.** The validation process at the lesion level remains insufficient. The majority of studies used image-level labels for explanation validation, but these labels do not prove that the highlighted areas match the actual disease locations [2], [10]–[12], [15].

- **Scarcity of clinical and user studies.** The field lacks sufficient research on clinical applications and user testing. The field lacks sufficient research on reader studies and clinical trials because most studies use small participant groups [1]–[3], [10], [12], [15].

2) Model and explanation design constraints.

- **Inconsistent and low-fidelity saliency maps.** The evaluation of different XAI methods produces distinct saliency maps which sometimes point to non-pathological areas or image noise [2], [10]–[12]. The explanations produced by some methods might appear valid, but they could lead to incorrect interpretations.
- **Over-reliance on coarse visual saliency.** Most research focuses on Grad-CAM heatmaps which show model attention points but do not explain the specific concepts or minimal changes that affect decisions [5], [7], [9], [22]–[25]. The field of DR needs more research on textual explanations and concept-based and counterfactual explanations.
- **Limited exploration of inherently interpretable models.** The development of inherently interpretable models has not received sufficient attention in research. The development of interpretable radiomics and prototype-based networks and attention mechanisms remains rare because most research focuses on applying post-hoc methods to conventional CNNs [3], [11], [15]–[18], [24].

3) Deployment, generalizability, and regulatory issues.

- **Dataset bias and generalizability.** Heavy reliance on a small number of public datasets with limited demographic and device diversity raises concerns about out-of-distribution performance and bias [2]–[4], [6], [10], [11], [16]–[18]. Only a fraction of studies evaluated models on external cohorts.
- **Computational cost and workflow integration.** The use of limited public datasets with restricted demographic and device diversity creates concerns about model performance outside of training data and potential bias [2]–[4], [6], [10], [11], [16]–[18]. The research includes only a few studies that tested their models on independent test sets.
- **Regulatory and ethical considerations.** The research lacks sufficient discussion about regulatory requirements and ethical standards which affect explanation system deployment. The research lacks sufficient discussion about post-market surveillance and ethical aspects of explanation systems because only a few studies addressed these topics (e.g. [10]). The research lacks answers about what minimum explanation accuracy standards need for regulatory approval and how to track explanation system behavior throughout time as shown in table 5 and illustrated in figure 9.

Table 5: Key Challenges and Research Gaps Identified

Challenge / Gap	Frequency Mentioned (n, %)
Lack of standardized evaluation metrics	38 (73.1%)
Limited clinical validation and user studies	46 (88.5%)
Inconsistency and low fidelity of saliency maps	22 (42.3%)
Over-reliance on visual saliency maps	18 (34.6%)
Insufficient lesion-level validation	41 (78.8%)
Dataset bias and generalizability concerns	29 (55.8%)
Computational cost and workflow integration	12 (23.1%)
Regulatory and ethical considerations	4 (7.7%)
Limited exploration of inherently interpretable models	15 (28.8%)
Lack of longitudinal and multi-modal integration	8 (15.4%)

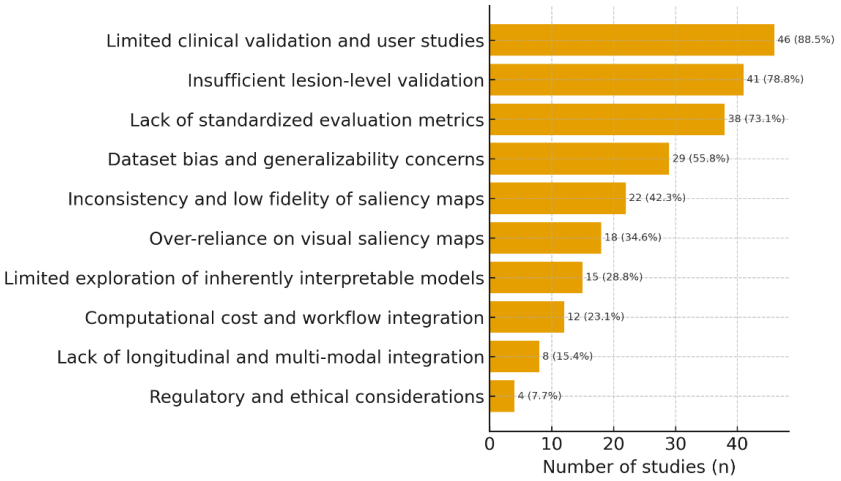

figure 9. depicts the frequency of reported challenges and research gaps in XAI of DR

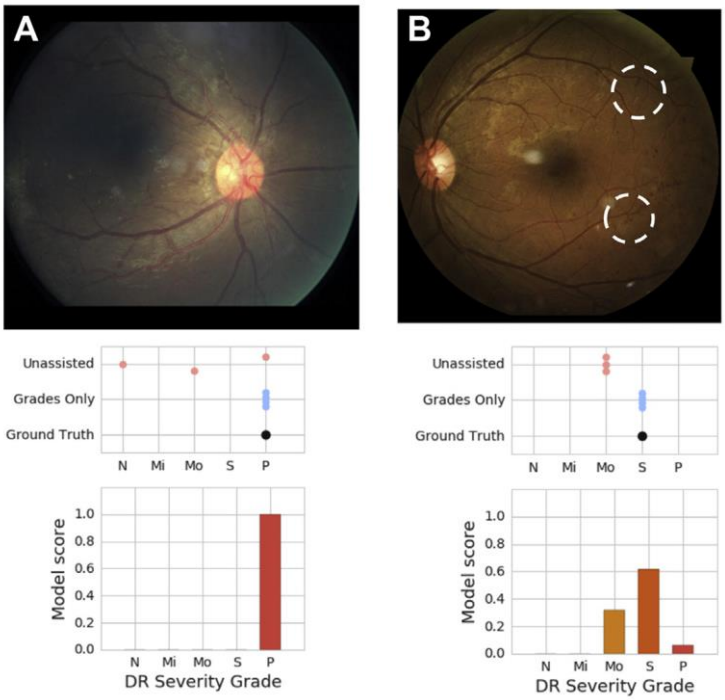

figure 10. improving diabetic retinopathy diagnosis [1]

Figure 10. Examples of cases where the model evaluates improved diabetic retinopathy diagnosis compared to readings without assistance. A, underdiagnosis of proliferative diabetic retinopathy without assistance. B, lack of diagnosis of severe non-proliferative diabetic retinopathy without assistance.

For each panel, the top part shows a fundus image; the middle part displays the diabetic retinopathy grades for readers in cases of no assistance (red dots) and grades only (blue dots) (3 for each arm), in addition to the reference standard grade (black dot), for each image [1].

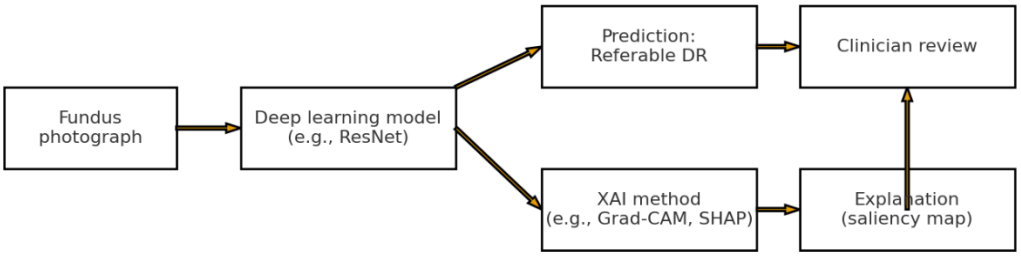

figure 10. conceptual workflow of explainable AI (XAI) of DR detection

Figure 10 depicts that the explanations can enhance trust, verification, and error detection by showing why the models made their decision. The QUADAS-2 and XAI-checklist assessments showed that patient selection and index test validation and XAI evaluation rigor suffered from high or unclear risk of bias which supports Gaps 1–3 [2], [3], [10], [11].

DISCUSSION

4.1 Principal Findings

The systematic review analyzed 52 studies about XAI in DR detection which shows a growing yet inconsistent research area. The majority of systems employ top-performing CNN architectures including ResNet and EfficientNet and VGG and Inception and DenseNet for DR detection while using post-hoc saliency and attribution methods like Grad-CAM and SHAP and LIME and Integrated Gradients [2]–[4], [5]–[9], [10]–[14], [16]–[20], [22]–[25]. The research field investigates two distinct approaches to model interpretation through radiomics-based and prototype-based and evidence-based and attention-driven architectures [3, [11], [15]–[18], [24].

The evaluation of explanations depends mainly on visual assessment but ECS and IoU and DSC and lesion-level metrics appear in less than half of the studies and ophthalmologist participation in user studies remains infrequent [1]–[3], [10]–[12], [15]. The quantitative results show that explanation methods require optimization to detect more than half of actual lesions, yet their medical trustworthiness remains unclear [2].

The performance of models remains unaffected by post-hoc XAI but user testing shows that explanation methods enhance human-AI teamwork by improving accuracy and confidence and reducing reading time [1]–[3], [11], [15]. The current state shows a wide difference between laboratory-based XAI demonstrations and actual medical practice because researchers lack proper external testing and fail to address regulatory requirements [2], [3], [10], [11].

4.2 Research Gaps

The research findings together with quality evaluation methods reveal multiple essential knowledge gaps.

1. **Standardized, multi-dimensional evaluation of explanations.** The medical field requires standardized evaluation methods which assess explanations through multiple dimensions. The medical field requires immediate development of standardized evaluation systems to assess explanation fidelity and plausibility and usefulness and robustness in DR screening. The evaluation system needs to integrate lesion-level metrics (ECS, IoU, DSC) with sensitivity analysis through degradation tests and clinician-based performance assessments for task completion with and without explanation [1]–[3], [10]–[12], [15].
2. **Prospective clinical validation and workflow integration.** The medical field requires both clinical testing of XAI systems in real-world screening operations and their integration into healthcare workflows. The current evidence base consists mainly of studies conducted after the fact. The implementation of XAI systems in real-world screening programs through prospective trials will show their effectiveness in clinical decision-making and their ability to decrease errors and gain clinician acceptance [1]–[3], [10], [11], [15]. The research needs to assess how XAI systems affect system performance and IT system compatibility and human operator workload.
3. **Beyond saliency maps: richer explanation paradigms.** The current explanation methods based on saliency maps need improvement through the development of additional explanation approaches. The current explanation methods based on heatmaps might restrict their potential impact on clinical practice. The development of explanation methods which use concepts and counterfactuals and examples will help ophthalmologists perform better lesion assessment and severity evaluation and progression prediction [7]–[9], [11], [20], [23]–[25]. The development of prototype-based and evidence-based models shows promise for further investigation according to [11], [15], [24].
4. **Generalizability, fairness, and regulation.** The field requires studies to establish system-wide applicability and fairness standards and regulatory frameworks. The evaluation of system performance needs to occur through external testing with different patient groups and medical equipment and healthcare environments to verify widespread applicability and detect potential biases [2]–[4], [6], [10], [11], [16]–[18]. The current standards for XAI implementation in medical devices remain insufficient for regulatory agencies to enforce transparency requirements and post-market surveillance [10]. The development of acceptable explanation behaviors and documentation standards requires joint efforts between AI developers and clinicians and regulators and ethicists.

4.3 Trends and Future Directions

The current trends in XAI for DR show the following developments:

- **Hybrid and ensemble XAI**, The combination of different explanation methods through hybrid and ensemble XAI systems produces multiple explanation perspectives (e.g. Grad-CAM for localization and SHAP for feature attribution) [2], [3], [7], [10], [13], [22], [23].
- **Concept-based and prototype-based models**, the prediction explanation system uses concept-based and prototype-based models to explain results through general retinal concepts and lesion patch similarities [11], [15], [24].
- **Counterfactual explanations**, the explanation system shows how small modifications in lesion amount or distribution would produce different prediction results [7], [9], [20], [23]–[25].
- **Attention-driven architectures**, the attention-driven architecture uses multi-head or hierarchical attention mechanisms to detect important clinical areas which produce more reliable explanations [16], [17].
- **Federated and privacy-preserving XAI**, The XAI system uses federated learning to analyze distributed medical data without exposing patient information to central point's [6], [16], [18].

The different trends in XAI development address particular weaknesses in DR systems through their implementation. The prototype-based and concept-based models solve the problem of saliency map accuracy (Gaps 4–5) while federated XAI addresses both dataset bias and patient data privacy (Gap 7).

4.4 Implications for Practice

The research indicates that XAI-enhanced DR systems function as useful additional tools for ophthalmologists and screening program leaders, but they should not replace human decision-making. The explanations in these systems help users understand specific areas and detect errors but users should treat them as helpful clues instead of absolute proof for lesion identification [1]–[3], [10]–[12], [15]. Users need to evaluate the connection between highlighted areas and actual medical conditions while keeping an eye out for potential false positives in saliency map results.

AI developers need to create XAI pipelines which focus on user needs and follow clinical standards. The evaluation process requires both detailed analysis of lesions and testing with human participants while designers must create user-friendly interfaces that show explanations without information overload and developers must perform regular external testing and surveillance [1]–[3], [10]–[12], [15].

The evaluation process for AI-enabled DR devices requires specific guidelines from regulators and policymakers about

explanation requirements and documentation needs and validation procedures and post-market monitoring [10]. The assessment of explainability requires both technical evaluation and measurement of its effects on safety outcomes and effectiveness and fairness.

4.5 Limitations of This Review

The evaluation contains specific restrictions which affect its results. The study only included English-language peer-reviewed articles which might produce biased results because they represent more developed research while excluding new preprint studies. The research focused on fundus photography so the results might not apply to OCT-based DR systems and other ophthalmic imaging technologies. The study required narrative synthesis because different datasets and architectures and XAI methods and reporting standards made meta-analysis impossible. The researchers faced challenges in assessing risk of bias because authors failed to provide sufficient details about patient selection and reference standards and explanation evaluation methods [2], [3], [10], [11].

CONCLUSION

The development of Explainable AI for DR detection started with basic heatmaps before advancing to complex attribution techniques and interpretable system designs. The DR AI literature now includes substantial research using CNN models with Grad-CAM-style saliency maps and SHAP/LIME/Integrated Gradients explanations which show potential for human-AI collaboration improvement [1]–[3], [5]–[9], [10]–[15], [22].

The current state of explanation evaluation lacks consistency and most studies fail to provide strong methodological foundations while lesion-level validation occurs rarely and large-scale prospective clinical trials are hard to find and deployment requirements such as computation time and bias management and regulatory compliance are just starting to be studied [2], [3], [10]–[12]. The development of standardized multi-dimensional evaluation frameworks together with XAI implementation in real-world screening operations and research into inherently interpretable models and regulatory framework compliance will lead to clinically reliable and regulatory-grade systems [1]–[3], [10], [11],[15].

The combination of explainable AI technology could connect advanced automated systems with dependable medical choices for DR screening operations. The achievement of this goal demands ongoing teamwork between AI scientists and medical doctors and human-computer interface experts and regulatory bodies and people who need medical care.

Funding: This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant No. KFU254123]

REFERENCES

1. R. Sayres *et al.*, “Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy,” *Ophthalmology*, vol. 126, no. 4, pp. 552–564, 2019, doi: 10.1016/j.ophtha.2018.11.016.
2. T. Van Craenendonck *et al.*, “Systematic Comparison of Heatmapping Techniques in Deep Learning in the Context of Diabetic Retinopathy Lesion Detection,” *Transl. Vis. Sci. Technol.*, vol. 9, no. 2, p. 64, 2020, doi: 10.1167/tvst.9.2.64.
3. T. Alghamdi, “Detection of Retinopathy Diabetic Using Explainable AI: Interpretable Deep Learning Models in Clinical Practice,” *J. Eng. Sci. Comput.*, vol. 2, no. 1, p. 018, 2025, doi: 10.63070/jesc.2025.018.
4. J. Civit-Masot, F. Luna-Perejón, L. Muñoz-Saavedra, *et al.*, “An Explainable Ensemble Based Approach to Diabetic Retinopathy Grading,” *Res. Sq.* [Preprint], 2025, doi: 10.21203/rs.3.rs-6878828/v1.
5. R. R. Selvaraju *et al.*, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626.
6. S. Kansal *et al.*, “RetinoDeep: Leveraging Deep Learning Models for Advanced Retinopathy Diagnostics,” *Sensors*, vol. 25, no. 16, p. 5019, 2025, doi: 10.3390/s25165019.
7. S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 4765–4774.
8. M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 1135–1144.
9. M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic Attribution for Deep Networks,” in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 3319–3328.
10. E. Shakeri *et al.*, “Using SHAP Analysis to Detect Areas Contributing to Diabetic Retinopathy Detection,” in *Proc. 2022 IEEE Int. Conf. Inf. Reuse Integr. Data Sci. (IRI)*, 2022, pp. 157–162, doi: 10.1109/IRI54793.2022.00046.
11. D. K. Kumar, A. Wong, and G. W. Taylor, “Discovery Radiomics With CLEAR-DR: Interpretable Computer Aided Diagnosis of Diabetic Retinopathy,” *IEEE Access*, vol. 7, pp. 25891–25908, 2019, doi: 10.1109/ACCESS.2019.2893635.
12. H. Kheradfallah *et al.*, “Annotation and Segmentation of Diabetic Retinopathy Lesions: An Explainable AI Application,” in *Proc. SPIE Med. Imaging 2022: Imaging Informatics for Healthcare, Research, and Applications*, vol. 12037, 2022, p. 120370L, doi: 10.1117/12.2612576.
13. E. N. Volkov and A. Averkin, “Hybrid Explainable Framework for Diabetic Retinopathy Classification from Fundus Images,” in *Proc. 2024 IEEE Int. Conf. Soft Comput. Meas. (SCM)*, 2024, pp. 554–557, doi: 10.1109/SCM62608.2024.10554254.
14. K. Mittal, “Deep Learning Model With Game Theory-Based Gradient Explanations for Retinal Images,” in *Proc. Int. Conf. Artif. Intell. Data Sci.*, 2023, pp. 211–224, doi: 10.1007/978-981-99-0609-3_15.

15. K. Djoumessi *et al.*, "An Inherently Interpretable AI Model Improves Screening Speed and Accuracy for Early Diabetic Retinopathy," *PLOS Digit. Health*, vol. 4, no. 5, p. e0000831, 2025, doi: 10.1371/journal.pdig.0000831.
16. Z. Wang *et al.*, "Diabetic Retinopathy Classification Using a Multi-Attention Residual Refinement Architecture," *Sci. Rep.*, vol. 15, p. 5269, 2025, doi: 10.1038/s41598-025-15269-1.
17. Ş. Y. Atcı, "An Integrated Deep Learning Approach for Computer-Aided Diagnosis of Diverse Diabetic Retinopathy Grading," in *Lect. Notes Comput. Sci.*, Springer, 2024, pp. 95–108, doi: 10.1007/978-3-031-52787-6_8.
18. [S. Bitto *et al.*, "Explainable AI Based Deep Ensemble Convolutional Learning for Multi-Categorical Ocular Disease Prediction," *EAI Endorsed Trans. Artif. Intell. Robot.*, vol. 4, p. e1, 2025, doi: 10.4108/airo.9234.
19. K. Duvvuri *et al.*, "Grad-CAM for Visualizing Diabetic Retinopathy," in *Proc. 2022 Int. Conf. Emerg. Techniques Comput. Intell. (ICETCI)*, 2022, pp. 87–91, doi: 10.1109/INCET54531.2022.9824598.
20. L. Li *et al.*, "Case Study: Explaining Diabetic Retinopathy Detection Deep CNNs via Integrated Gradients," *arXiv* preprint arXiv:1709.09842, 2017.
21. U.S. Food and Drug Administration, *Artificial Intelligence and Machine Learning (AI/ML) Software as a Medical Device Action Plan*, Jan. 2021, accessed Nov. 15, 2025.
22. Z. Wang and J. Yang, "Diabetic Retinopathy Detection via Deep Convolutional Networks for Discriminative Localization and Visual Explanation," *arXiv* preprint arXiv:1703.10757, 2017.
23. Q.-T. Dao, H.-Q. Trinh, and V.-A. Nguyen, "An Effective and Comprehensible Method to Detect and Evaluate Retinal Damage Due to Diabetes Complications," *PeerJ Comput. Sci.*, vol. 9, p. e1585, 2023, doi: 10.7717/peerj-cs.1585.
24. S. Hesse, "INSightR-Net: Interpretable Neural Network for Regression Using Similarity-Based Comparisons to Prototypical Examples," in *Lect. Notes Comput. Sci.*, vol. 13435, Springer, 2022, pp. 502–512, doi: 10.1007/978-3-031-16437-8_48.
25. L. Listyalina *et al.*, "Fovea and Diabetic Retinopathy: Understanding the Relationship Using a Deep Interpretable Classifier," *Comput. Methods Programs Biomed. Update*, vol. 2, p. 100059, 2022, doi: 10.1016/j.cmpbup.2022.100059.