Artificial intelligence driven prognostic models for predicting diabetic retinopathy risk in type 2 diabetes mellitus: A systematic review

Muhammad Muaz Shahriman Teruna, MPH[1], Tajul Rosli Razak, PhD[2], Siti Munira Yasin, DrPH[1], Abdullah Ashraf Rafique Ali, MOPH[3], Muhammad Muzzammil Mohamad Salleh, MPH[1], Mohamad Zuhair Mohamed Yusoff, MPH[1], Muhammad Hariz Ammar Khebir, MPH[1], Nur Adila Che Rameli, MPH[1], Muhammad Irfan Mohd Sallehhudin, MPH[1], Muhamad Syazni Mohamad Asraff, MPH[1], Mohamad Rodi Isa, DrPH[1]

[1]Department of Public Health Medicine, Faculty of Medicine, Universiti Teknologi MARA, Sungai Buloh, Selangor, Malaysia, [2]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia, [3]Department of Ophthalmology, Faculty of Medicine, Universiti Teknologi MARA, Sungai Buloh, Selangor, Malaysia

ABSTRACT

Introduction: Diabetic retinopathy remains a leading cause of preventable blindness among individuals with type 2 diabetes mellitus and is driven by the increasing global burden of diabetes. Early identification of individuals at high risk is essential for timely intervention. In recent years, artificial intelligence, including machine learning and deep learning, has emerged as a promising approach for prognostic modelling. However, evidence on artificial intelligence–driven models for predicting diabetic retinopathy risk remains fragmented. This systematic review aims to synthesise current evidence on artificial intelligence–driven prognostic models for predicting the risk of diabetic retinopathy among individuals with type 2 diabetes mellitus, focusing on model characteristics, predictor variables, performance, and validation strategies.

Materials and Methods: A comprehensive search of PubMed, Scopus, Web of Science, and ScienceDirect was performed for studies published between January 2016 and December 2025. Eligible studies included those applying artificial intelligence–based models for diabetic retinopathy risk prediction in adult populations with type 2 diabetes mellitus and reporting model performance metrics. Data were extracted and synthesised narratively, and methodological quality was assessed using the Newcastle–Ottawa Scale.

Results: A total of 1040 records were identified, with eight studies included after screening. A wide range of artificial intelligence algorithms was applied, with ensemble models such as extreme gradient boosting and random forest demonstrating superior performance. Key predictors consistently included glycated haemoglobin levels, duration of diabetes, blood pressure, lipid profile, and renal function markers, alongside emerging metabolomic biomarkers. Model performance ranged from moderate to excellent, with area under the receiver operating characteristic curve values between 0.68 and 0.97. Most studies employed internal validation techniques such as cross-validation or training and testing data splits, while external validation was largely absent. Overall methodological quality was high, although variability in study design, predictor selection, and reporting was observed.

Conclusion: Artificial intelligence–driven prognostic models show substantial potential in predicting diabetic retinopathy risk among individuals with type 2 diabetes mellitus, particularly through the integration of clinical and high-dimensional data. Despite promising predictive performance, limitations in external validation, heterogeneity, and reporting of clinical utility hinder translation into practice. Future research should prioritise external validation, standardised reporting frameworks, and clinically interpretable models to enhance applicability in real-world healthcare settings.

KEYWORDS:
Diabetic Retinopathy, Type 2 Diabetes Mellitus, Artificial Intelligence, Machine Learning, Prognostic Model

INTRODUCTION

Diabetes mellitus (DM) represents a major global health challenge, with an estimated 463 million people affected worldwide in 2019 and projections indicating a rise to 700 million by 2045.[1] Type 2 diabetes mellitus (T2DM), characterized by insulin resistance alongside varying degrees of β-cell dysfunction, accounts for approximately 90% of all diabetes cases.[2] Among the most devastating microvascular complications of T2DM, diabetic retinopathy (DR) remains a leading cause of preventable blindness in working-age adults globally.[3] The continuously climbing prevalence of T2DM has led to a 67% increase in vision impairment and a 27% increase in blindness attributable to DR from 1990 to 2010.[4] Early detection and timely intervention are therefore paramount in preventing the progression of DR to sight-threatening stages, including severe non-proliferative diabetic retinopathy (NPDR) and proliferative diabetic retinopathy (PDR).

The pathogenesis of DR involves multiple factors, including chronic hyperglycaemia, angiogenic factors, oxidative stress,

This article was accepted: 10 July 2026
Corresponding Author: Mohamad Rodi Isa
Email: rodi@uitm.edu.my

hypertension, and hyperlipidaemia.[5] As DR progresses, the development of new vessels leads to vitreous haemorrhage and tractional membrane formation, which can result in devastating deterioration of vision.[6] DR is classified according to the International Classification of DR scale into stages: no DR, mild NPDR, moderate NPDR, severe NPDR, and PDR, with sight-threatening DR (STDR) defined as the combination of severe NPDR and PDR.[7]

The emergence of artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), has introduced transformative possibilities in diabetology and ophthalmology.[8] AI is broadly defined as technology that enables computers and machines to simulate human intelligence and solve complex problems, encompassing subdomains such as ML, programs that improve over time with experience and DL, which uses artificial neural networks and large datasets to tackle computationally complex problems.[9] Within the field of endocrinology and diabetes, DR is among the conditions for which there is the largest body of evidence for the use of AI technologies in disease detection, prediction, and risk assessment.[10]

While substantial research has focused on AI-based diagnostic systems for DR primarily through automated retinal image analysis, there is a growing body of literature on AI-driven prognostic models that aim to predict the risk of developing DR or its progression in patients with T2DM. These prognostic models leverage clinical, demographic, biochemical, genetic, and imaging data to stratify patients by risk, enabling targeted preventive interventions.[11] This systematic review aims to comprehensively synthesize the current evidence on AI-driven prognostic models for predicting the risk of DR in individuals with T2DM, examining the methodologies employed, and the performance metrics achieved.

MATERIALS AND METHODS

This systematic review was conducted following the 2020 Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.[12] This review was registered on the PROSPERO platform (CRD420251271494).

Search strategy

A comprehensive literature search was conducted in PubMed, EMBASE, Scopus, Web of Science, and ScienceDirect between 5 January 2026 and 30 January 2026 to identify relevant studies published between 1 January 2016 and 31 December 2025. The search strategy was developed using a combination of controlled vocabulary, including Medical Subject Headings (MeSH) in PubMed and Emtree terms in EMBASE, together with free-text terms related to DR, T2DM, AI, ML, DL, prognostic modelling, risk prediction, and predictive analytics.

To maximise the sensitivity and comprehensiveness of the search, various synonyms, spelling variations, truncation symbols, and Boolean operators were incorporated. Terms related to DR included "Diabetic Retinopathy," relevant MeSH and Emtree terms, and retinopathy-related keywords. For the diabetes component, terms such as "Diabetes Mellitus, Type 2," "type 2 diabetes mellitus," "type 2 diabetes," T2DM, and diabetes were used. The AI domain included terms such as "Artificial Intelligence," AI, "machine learning," "deep learning," "neural networks," and "artificial neural networks." To capture studies focused on prediction and prognosis, additional terms included "prognosis," "prognostic modelling," "prediction," "risk prediction," "risk assessment," "risk stratification," "prediction models," "prognostic models," "clinical prediction models," "risk scores," and "prediction algorithms."

These concepts were combined using Boolean operators to identify studies that examined the application of AI and ML techniques for predicting the risk or progression of DR among individuals with T2DM. Search strategies were adapted for each database according to their specific indexing systems and search functionalities, including the use of MeSH terms in PubMed and Emtree terms in EMBASE, to ensure optimal retrieval of relevant records. The specific strategy applied for each database is detailed in table I.

Eligibility criteria

This systematic review included original studies published between January 2016 and December 2025 that evaluated AI-driven prognostic or risk prediction models DR among individuals with T2DM. Eligible studies involved adult populations (≥18 years) with T2DM and assessed the development or progression of DR as an outcome. Studies were required to apply AI or ML techniques for risk prediction, prognostic modelling, or stratification, and to report at least one model performance metric such as area under the receiver operating characteristic (AUROC) curve, sensitivity, specificity, calibration, or related indicators. Observational study designs (cohort, case-control, or cross-sectional) and methodological/model development studies were included, provided they were published in peer-reviewed journals and written in English.

Studies were excluded if they did not focus specifically on T2DM populations or did not evaluate DR as a primary or clearly defined outcome. Articles that used AI solely for image-based detection, screening, or classification of existing DR without a prognostic or risk prediction component were excluded, as were studies lacking sufficient methodological detail or reporting no model performance metrics. Non-original articles such as reviews, systematic reviews, meta-analyses, editorials, conference abstracts without full texts, case reports, and letters were also excluded. In addition, studies involving animal models, paediatric populations, gestational diabetes, or type 1 diabetes mellitus were not considered, as well as publications outside the specified date range or not available in English.

Study selection

Following the database searches, all retrieved records were imported into reference management software (e.g. Endnote) and duplicates were removed. Two independent reviewers (M.M.S.T. and T.R.R.) screened titles and abstracts against eligibility criteria. Full text articles of potentially relevant studies were retrieved and assessed for final inclusion. Disagreements were resolved through discussion or consultation with a third reviewer (M.R.I.) when consensus could not be reached.

Data extraction

Data extraction was conducted using a standardized, pilot-tested form (e.g., Endnote) to ensure consistency and completeness across included studies. Two reviewers (M.M.S.T. and T.R.R.) independently extracted key information, including study characteristics (first author, publication year, country, study design, and sample size), population details, and outcome definitions related to DR. Detailed information on the AI models was also collected, including the type of algorithm used, input variables or predictors, model development approach, and validation methods (internal and/or external validation). In addition, performance metrics such as AUROC curve, sensitivity, specificity, accuracy, calibration measures, and other relevant indicators were extracted.

Risk of bias assessment

The methodological quality and risk of bias of the included studies were assessed using the Prediction Model Risk of Bias Assessment Tool (PROBAST). PROBAST was selected because it is a validated and widely recommended instrument specifically developed for evaluating diagnostic and prognostic prediction model studies, including those incorporating AI and ML approaches.[13] Unlike generic quality assessment tools designed for observational studies, PROBAST provides a structured framework to assess potential sources of bias that may affect the development, validation, and performance of prediction models. The tool evaluates four key domains: participants, predictors, outcome, and analysis, encompassing a total of 20 signalling questions. Each domain was rated as having low, high, or unclear risk of bias, and an overall risk of bias judgement was subsequently assigned for each study. Two reviewers (M.M.S.T. and T.R.R.) independently performed the assessment, and any disagreements were resolved through discussion and consensus with a third reviewer (M.R.I).

Data synthesis

A narrative synthesis approach was undertaken due to the anticipated heterogeneity in study designs, AI algorithms, predictor variables, outcome definitions, and reporting of performance metrics across the included studies. Extracted data were systematically organised into structured tables and synthesised according to key domains, including study characteristics, types of AI models, predictor variables, and model performance. The findings were analysed thematically to identify common patterns in model development approaches, frequently used predictors, and variations in predictive performance (e.g., discrimination and calibration). Comparisons were made across studies to highlight differences in algorithm performance, validation strategies (internal versus external), and clinical applicability.

RESULTS

Study selection and characteristics

A total of 1040 studies were retrieved from the databases and screened based on predefined criteria. After removing ineligible records, 537 studies remained for title and abstract screening. A total of 61 studies progressed to full-text screening. Studies were further evaluated for relevance, methodological quality, and outcome reporting during this phase. After final appraisal, eight studies were included in this review (Figure 1).

The included eight studies all published between 2016 and 2025, reflecting the growing application of AI in prognostic DR. The studies were predominantly conducted in Asia and Europe, with a mix of retrospective cohort, cross-sectional, and secondary data analyses. Sample sizes varied substantially, ranging from a few hundred participants to large-scale datasets exceeding 30,000 individuals, indicating variability in study power and generalisability. The findings of the included studies were systematically synthesised and summarised into four key domains namely AI model characteristics, predictor variables, model performance, and validation strategies, to facilitate structured comparison and critical interpretation of the evidence. The key characteristics of these studies are outlined in table II.

AI model characteristics

Across the eight included studies, a wide range of AI and ML algorithms were applied, with a clear dominance of ensemble-based methods, particularly extreme gradient boosting (XGBoost). Multiple studies consistently demonstrated the use of advanced algorithms such as XGBoost, random forest (RF), support vector machine (SVM), artificial neural networks (ANN), and gradient boosting models, reflecting a shift from traditional statistical models toward more flexible, data-driven approaches.[15-16,19] Notably, explainable artificial intelligence (XAI) techniques, including Explainable Boosting Machine (EBM) and Shapley Additive exPlanations (SHAP), were increasingly integrated to enhance interpretability and clinical relevance of model outputs.[14,19] These approaches allowed not only prediction but also identification of key contributing features, facilitate clinical decision-making in routine screening and public health programmes.[22]

Predictor variables

The predictor variables across studies showed a consistent pattern, combining traditional clinical risk factors with emerging biochemical and metabolomic markers. Classical predictors such as duration of diabetes, HbA1c, blood pressure, lipid profile, and body mass index were frequently identified as important determinants of DR risk.[15,17-18] Some studies quantified these associations using effect estimates, for example, insulin use (OR 3.561) and increasing duration of diabetes (OR increase 9.3% per year).16 In addition, newer studies incorporated advanced predictors such as renal function markers (eGFR, creatinine), lipid-related indicators, and metabolomic biomarkers (e.g., tryptophan, tyrosine), reflecting the increasing use of high-dimensional data in AI modelling.[14,18] However, most studies relied on feature importance rankings rather than conventional epidemiological measures such as adjusted hazard ratios, limiting direct interpretability.

Model performance

Model performance varied across studies but generally demonstrated moderate to excellent predictive ability. Ensemble learning methods, particularly XGBoost, consistently achieved superior discrimination, with AUROC values ranging from 0.68 to 0.97 depending on dataset size

Table I: Primary search strings used in each database

Database	Search Strategy
PubMed	("Diabetic Retinopathy"[MeSH] OR "diabetic retinopathy" OR retinopath*) AND ("Diabetes Mellitus, Type 2"[MeSH] OR "type 2 diabetes mellitus" OR "type 2 diabetes" OR T2DM) AND ("Artificial Intelligence"[MeSH] OR "artificial intelligence" OR AI OR "machine learning" OR "deep learning" OR "neural network*" OR "artificial neural network*") AND ("Prognosis"[MeSH] OR prognos* OR predict* OR "risk prediction" OR "risk assessment" OR "risk stratification" OR "prediction model*" OR "prognostic model*" OR "clinical prediction model*" OR "risk score*" OR "prediction algorithm*")
EMBASE	('diabetic retinopathy'/exp OR 'diabetic retinopathy' OR retinopath*) AND ('type 2 diabetes mellitus'/exp OR 'type 2 diabetes mellitus' OR 'type 2 diabetes' OR T2DM) AND ('artificial intelligence'/exp OR 'artificial intelligence' OR 'machine learning'/exp OR 'machine learning' OR 'deep learning' OR 'neural network*' OR 'artificial neural network*') AND ('prognosis'/exp OR prognos* OR predict* OR 'risk prediction' OR 'risk assessment' OR 'risk stratification' OR 'prediction model*' OR 'prognostic model*' OR 'clinical prediction model*' OR 'risk score*' OR 'prediction algorithm*')
Scopus	(("diabetic retinopathy" OR retinopath*) AND ("type 2 diabetes mellitus" OR "type 2 diabetes" OR T2DM) AND ("artificial intelligence" OR AI OR "machine learning" OR "deep learning" OR "neural network*" OR "artificial neural network*") AND (prognos* OR predict* OR "risk prediction" OR "risk assessment" OR "risk stratification" OR "prediction model*" OR "prognostic model*" OR "clinical prediction model*" OR "risk score*" OR "prediction algorithm*"))
Web of Science	(("diabetic retinopathy" OR retinopath*) AND ("type 2 diabetes mellitus" OR "type 2 diabetes" OR T2DM) AND ("artificial intelligence" OR AI OR "machine learning" OR "deep learning" OR "neural network*" OR "artificial neural network*") AND (prognos* OR predict* OR "risk prediction" OR "risk assessment" OR "risk stratification" OR "prediction model*" OR "prognostic model*" OR "clinical prediction model*" OR "risk score*" OR "prediction algorithm*"))
ScienceDirect	("diabetic retinopathy" OR retinopath*) AND ("type 2 diabetes mellitus" OR "type 2 diabetes" OR T2DM) AND ("artificial intelligence" OR AI OR "machine learning" OR "deep learning" OR "neural network*" OR "artificial neural network*") AND (prognos* OR predict* OR "risk prediction" OR "risk assessment" OR "risk stratification" OR "prediction model*" OR "prognostic model*" OR "clinical prediction model*" OR "risk score*" OR "prediction algorithm*")

and model complexity.[14,17,19] High-performing models also reported strong accuracy, sensitivity, and specificity, with one study achieving an AUROC of 0.97 alongside balanced precision and recall metrics.[14] Larger cohort-based studies demonstrated robust performance with AUROC values above 0.80, highlighting the advantage of large datasets in improving model generalisability.[18] In contrast, studies with smaller sample sizes or limited predictors reported more modest performance, indicating variability influenced by study design and data quality.[15]

Validation strategies

Validation approaches across the included studies were predominantly internal, with limited use of external validation. Most studies employed techniques such as k-fold cross-validation, train–test split, or three-way data partitioning to assess model performance and reduce overfitting.[14-16] Some cohort studies incorporated temporal validation using follow-up data, which provided a stronger assessment of predictive performance over time.[18] However, external validation using independent datasets was largely absent, representing a significant methodological limitation. Additionally, while internal validation was commonly performed, calibration and clinical utility assessments were infrequently reported, limiting the evaluation of real-world applicability of these models.[19]

Risk of bias assessment

Overall, the risk of bias assessment using the PROBAST tool indicated that six of the eight studies have a high overall risk of bias, primarily due to concerns within the analysis domain (Table III). Common methodological limitations included reliance on internal validation only, inadequate assessment of model calibration, potential overfitting arising from small sample sizes relative to model complexity, and insufficient reporting of model development procedures. Two studies were judged to have a low overall risk of bias because they employed large cohort datasets and incorporated stronger validation approaches, including temporal or external validation. Across all studies, the participants, predictors, and outcome domains were generally assessed as having low risk of bias, reflecting appropriate population selection, clearly defined predictors, and well-established DR outcomes. Applicability concerns were considered low for all included studies because the study populations, predictors, and outcomes were consistent with the review objective of evaluating AI-driven prognostic models for predicting DR risk among adults with T2DM.

Certainty of evidence assessment

The certainty of evidence was assessed using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach adapted for prognostic model research. The assessment considered five key domains, namely risk of bias, consistency, directness, precision, and publication bias, while also taking into account methodological heterogeneity and the extent of external validation reported across studies. Evidence was downgraded when serious concerns were identified, particularly regarding risk of bias and heterogeneity. Overall, the certainty of evidence was judged to be moderate (Table IV). Although the included studies consistently demonstrated that AI–driven models, particularly ensemble learning approaches such as XGBoost, RF, and gradient boosting methods, achieved moderate to excellent predictive performance for DR risk prediction among individuals with T2DM, several limitations reduced confidence in the overall evidence.

Table II: Characteristics of included studies

No.	Authors (Year)	Country	Study Design	Sample Size	AI Model Characteristics	Predictor Variables	Model Performance	Validation Strategies	Conclusion
1	Yagin et al. (2023)14	Turkey	Retrospective cohort	317	XGBoost, NGBoost, Explainable Boosting Machine (EBM) with Bayesian optimisation	Key biomarkers: Trp, Tyr, C16, DMA	Accuracy 91.25%, Precision 89.33%, Recall 91.24%, F1-score 89.37%, AUROC 0.97	10-fold cross-validation, internal validation	XAI-based models demonstrated excellent predictive accuracy and identified key metabolomic biomarkers for DR stratification.
2	Kotsiliti et al. (2017)15	UK	Retrospective cohort	6375	Logistic regression (LASSO), Random Forest, Gradient Boosting	Duration of diabetes, HbA1c, ACR, age	AUROC 0.73–0.77 (best: GBM)	10-fold cross-validation	Clinical and biochemical variables can reasonably predict DR risk, with improved performance using longitudinal data.
3	Tsao et al. (2018)16	Taiwan	Cross-sectional	3539	SVM, Decision Tree, ANN, Logistic Regression	Insulin, duration of diabetes,	Accuracy 79.5%, AUROC 0.839	Train-test split, three-way split validation	ML models effectively identify high-risk individuals, with interpretable predictors supporting clinical decision-making.
4	Rosu et al. (2024)17	Romania	Cross-sectional	377	Random Forest, XGBoost, SVM	SBP, LDL, BMI	AUROC: RF 0.62, XGBoost 0.68 (improved to 0.72 after tuning)	Internal validation (model tuning)	Cardiovascular risk factors play a significant role in DR progression, though predictive performance was moderate.
5	Zhao et al. (2022)18	China	Retrospective cohort	7943	XGBoost and other ML models	SUA, LDL-C, TC, eGFR, TG, HbA1c	AUROC 0.803, Accuracy 88.9%, Sensitivity 74.0%, Specificity 81.1%	Train-test split, temporal validation	XGBoost model demonstrated strong predictive performance and enabled earlier identification of high-risk patients.
6	Li et al. (2021)19	China	Retrospective cohort	32,452	XGBoost, Logistic Regression, RF, SVM with SHAP	HbA1c >8%, nephropathy, creatinine >100 μmol/L, insulin use	AUROC 0.90 (XGBoost best)	Internal validation, SHAP interpretability	XGBoost achieved high predictive accuracy, with SHAP enhancing clinical interpretability of risk factors.
7	Jiang & Li (2024)20	China	Retrospective cohort	349	Decision Tree, K-Nearest Neighbours, Logistic Regression Model, RF, SVM, and XGBoost.	Disease duration, diabetic peripheral neuropathy, insulin dosage, urinary protein, albumin.	Best model: Logistic regression. Validation set: AUROC 0.894 (95% CI 0.833–0.955), Recall 0.92, Accuracy 0.781, F1-score 0.800.	Random 7:3 train-validation split; LASSO feature selection; 1,000-bootstrap internal validation.	Logistic regression demonstrated superior predictive performance compared with other ML algorithms.
8	Kim et al. (2025)21	South Korea	Retrospective cohort	14,694	RF, XGBoost, AdaBoost, CatBoost.	Dyslipidaemia, cancer, hypertension, chronic kidney disease, HbA1c variability, age, BMI.	Best model: XGBoost. AUROC: 0.824, Accuracy: 75.13%, Sensitivity: 71.00%, Specificity: 75.23%.	Hyperparameter tuning; Synthetic Minority Oversampling Technique (SMOTE); 10-fold stratified cross-validation.	XGBoost demonstrated the best predictive performance for DR within 3 years among patients with T2DM.

Notes: AI = artificial intelligence, AUROC = area under the receiver operating characteristic, ANN = artificial neural network, ACR = albumin-to-creatinine ratio, BMI = body mass index, BP = blood pressure, C16 = hexadecanoyl carnitine, DMA = dimethylarginine, DR = diabetic retinopathy, EBM = explainable boosting machine, eGFR = estimated glomerular filtration rate, HDL = high-density lipoprotein, LDL = low-density lipoprotein, ML = machine learning, NGBoost = natural gradient boosting, OR = odds ratio, RF = random forest, SHAP = Shapley Additive exPlanations, SBP = systolic blood pressure, SVM = support vector machine, TC = total cholesterol, TG = triglycerides, Trp = tryptophan, Tyr = tyrosine, XAI = explainable artificial intelligence

Table III: Summary of reporting bias of included studies using PROBAST

No.	Authors (Year)	Participants	Predictors	Outcome	Analysis	Overall Risk of Bias	Applicability Concerns
1	Yagin et al. (2023)[14]	Low	Low	Low	High	High	Low
2	Kotsiliti et al. (2017)[15]	Low	Low	Low	High	High	Low
3	Tsao et al. (2018)[16]	Low	Low	Low	High	High	Low
4	Rosu et al. (2024)[17]	Low	Low	Low	High	High	Low
5	Zhao et al. (2022)[18]	Low	Low	Low	Low	Low	Low
6	Li et al. (2021)[19]	Low	Low	Low	High	High	Low
7	Jiang & Li (2024)[20]	Low	Low	Low	High	High	Low
8	Kim et al. (2025)[21]	Low	Low	Low	Low	Low	Low

Table IV: Summary of certainty of evidence assessment

Domain	Assessment	Justification
Risk of Bias	Serious	Several studies demonstrated methodological limitations, particularly in the analysis domain, including reliance on internal validation, inadequate calibration assessment, and potential overfitting.
Consistency	Not Serious	Most studies consistently reported moderate to excellent predictive performance, with artificial intelligence models demonstrating acceptable to high discrimination ability.
Directness	Not Serious	All included studies directly evaluated artificial intelligence–driven prognostic models for predicting DR risk among adults with type 2 diabetes mellitus.
Precision	Not Serious	Most studies included adequate sample sizes and reported performance metrics with acceptable precision.
Publication Bias	Undetected	No clear evidence of publication bias was identified; however, formal assessment was not feasible due to the limited number of studies.
Heterogeneity	Serious	Substantial heterogeneity was observed in study populations, predictor variables, artificial intelligence algorithms, outcome definitions, and validation approaches.
External Validation	Serious	Only a minority of studies performed temporal or external validation, limiting confidence in the generalisability of the models.
Overall Certainty of Evidence	Moderate	The evidence consistently supports the potential of artificial intelligence–driven prognostic models for predicting DR risk; however, methodological limitations, heterogeneity, and limited external validation reduced the overall certainty of evidence.

Most studies were retrospective in design and relied predominantly on internal validation techniques, with only a limited number incorporating temporal or external validation. Considerable heterogeneity was observed across studies with respect to study populations, predictor variables, AI algorithms, outcome definitions, and reported performance metrics. Furthermore, several studies were judged to have a high risk of bias in the analysis domain based on the PROBAST assessment, particularly due to concerns regarding overfitting, calibration assessment, and insufficient reporting of model development procedures. Nevertheless, the consistency of findings across different settings and populations, together with generally favourable discrimination performance, supports a moderate level of confidence that AI–driven prognostic models can effectively predict DR risk among adults with T2DM. Further large-scale prospective studies with external validation are required to increase the certainty of the evidence and support wider clinical implementation.

DISCUSSION

This systematic review synthesises the findings of the present studies on AI-driven prognostic models for DR among individuals with T2DM, with a focus on interpreting their methodological, predictive performance, and clinical relevance. The included studies demonstrate a growing shift toward the use of advanced ML algorithms and high-dimensional data to improve early risk stratification of DR. However, despite promising performance metrics, there remains considerable heterogeneity in model development, predictor selection, and validation approaches. This section critically integrates the evidence across the key domains of AI model characteristics, predictor variables, model performance, and validation strategies, highlighting both the strengths and limitations of current approaches while situating the findings within the broader context of translational applicability and public health impact.

AI model characteristics

The landscape of AI models applied to DR prediction and prognosis encompasses a broad spectrum of algorithmic approaches. ML algorithms, including logistic regression, decision trees, naïve Bayes classifiers, random forests, SVMs, and gradient boosting machines, have been widely employed for structured clinical data analysis.[23] DL models, particularly convolutional neural networks (CNNs), represent the dominant approach for image-based DR detection and progression prediction.[24] Specific architectures that have been applied include Inception-v3, ResNet, VGG-16, VGG-19, and deep neural networks (DNNs).[25] Transfer learning, whereby models pre-trained on large natural image datasets are fine-tuned on medical images, has been a particularly effective strategy for overcoming limited training data availability.[26]

Previous study demonstrated that DNN models consistently outperformed five traditional ML models such as logistic regression, decision tree, naïve Bayes, RF, and SVM across all

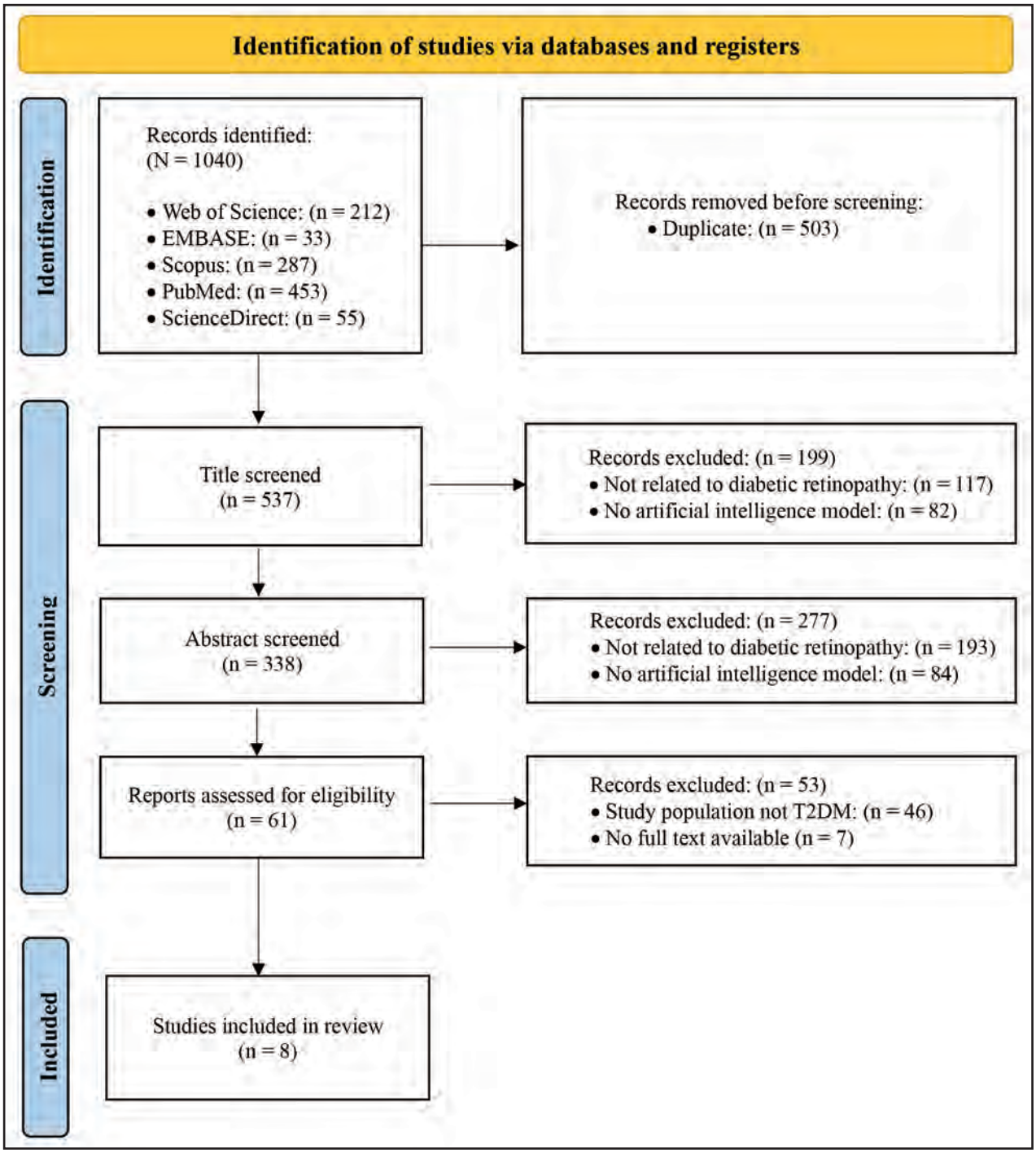

Fig. 1: Flowchart of the study selection process

evaluation metrics for predicting DR in patients with T2DM.[25] More recently, code-free DL (CFDL) or automated ML (AutoML) platforms have emerged as innovative tools that allow clinicians with limited coding experience to prototype novel clinical AI applications, demonstrating promising potential in DR care, including for DR progression prediction over time from ultra-wide-field images.27 Most of the study designs underlying AI model development for DR prediction are retrospective. The predominance of retrospective designs raises concerns about the generalizability of findings, as prospective validation in real-world clinical settings remains limited.[28]

Predictor variables

A consistent set of clinical and demographic predictor variables has been identified across multiple studies as being significantly associated with DR risk and progression. HbA1c has consistently emerged as one of the most significant predictors of DR. Previous study employed SHAP SHapley Additive exPlanations (SHAP) analysis to interpret their DL

model's predictions and found that HbA1c was one of the most significant predictors, with SHAP values validated against clinical reference standards for feature importance in DR.[25] Another study done found that HbA1c was significantly associated with DR, reinforcing its central role in prediction models.[29] Furthermore. the AI model developed in another study in 2025 similarly identified higher HbA1c levels as a risk factor for DR.[30]

Diabetes duration is another critical predictor, with longer duration consistently associated with increased DR risk because it reflects cumulative exposure to chronic hyperglycaemia. Persistent hyperglycaemia over time activates multiple pathogenic pathways, including the polyol pathway, formation of advanced glycation end-products (AGEs), activation of protein kinase C (PKC), and oxidative stress, which collectively contribute to retinal microvascular damage and endothelial dysfunction.[31] These mechanisms lead to structural alterations such as thickening of the capillary basement membrane and breakdown of the blood–retinal barrier, ultimately resulting in retinal ischemia and microaneurysm formation thereby increasing the likelihood of DR onset and progression.[32]

AI-specific methodological biases

In addition to conventional sources of bias encountered in prognostic model research, several AI-specific methodological biases were evident across the included studies and may have influenced the reported predictive performance. First, the widespread reliance on retrospective datasets and internal validation approaches increases the risk of model overfitting, whereby algorithms learn dataset-specific patterns that may not generalise to external populations. Second, many studies employed complex ML algorithms with limited transparency regarding feature selection, hyperparameter optimisation, and model tuning procedures, potentially introducing reproducibility and reporting bias. Third, algorithmic bias may arise from imbalances in patient characteristics, disease prevalence, ethnicity, healthcare utilisation patterns, or predictor distributions within the development datasets, resulting in systematic differences in model performance across populations.

Furthermore, several studies prioritised discrimination metrics such as AUROC and accuracy while providing limited assessment of calibration, model stability, and fairness, creating the risk of overly optimistic interpretations of predictive performance. The heterogeneity of predictor selection, outcome definitions, and data preprocessing methods across studies also introduces variability that may affect model transportability and external validity. Importantly, the limited use of explainable AI techniques restricts understanding of the underlying decision-making processes of complex algorithms, thereby reducing clinician trust and interpretability.[33] Collectively, these AI-specific methodological challenges highlight the need for greater adherence to emerging reporting and evaluation frameworks to improve transparency, reproducibility, fairness, and generalisability of future AI–driven prognostic models for DR prediction.

Model performance

The performance of AI models for DR prediction is typically evaluated using several standard metrics. The AUROC is the most commonly reported discriminative metric.34 Additional metrics include sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, F1 score, and for continuous outcomes, mean absolute error (MAE) and R^2.[7] For instance, recently published study reported that for all microvascular complications combined, predictive models achieved a mean c-statistic of 0.79 on internal validation and 0.72 on external validation.[35] Notably, DR models showed more modest performance compared to diabetic kidney disease models, which had the highest discrimination with c-statistics.

Apart from that, study done in China demonstrated that their DNN model outperformed traditional ML models for DR prediction, achieving accuracy of 0.778, F1 score of 0.776, and AUROC of 0.833, compared to logistic regression (accuracy 0.753, AUROC 0.822), decision tree (accuracy 0.630, AUROC 0.631), naïve Bayes (accuracy 0.718, AUROC 0.769), random forest (accuracy 0.758, AUROC 0.829), and SVM (accuracy 0.776, AUROC 0.831).25 Furthermore, another study done reported that Gaussian naïve Bayesian models accurately predicted diabetic complications including DR with an average AUROC of 0.86, while their expert-selected model achieved an average AUROC of 0.84 with 21 curated variables.[23] A critical concern in AI model performance is the variability observed across different populations and settings.

Validation strategies

Internal validation methods are essential for assessing model performance during development and for mitigating overfitting. The most commonly employed internal validation strategies include split-sample (hold-out) validation, cross-validation (k-fold, nested), bootstrapping, and temporal validation.[36] For example, internal validation methods was classified into into split-sample, bootstrapping, cross-validation, and temporal validation, noting that studies using temporal validation were categorized as internal validation regardless of how the studies labelled them.[34] Cross-validation, particularly k-fold cross-validation, has been widely adopted. For instance, a study done in China, the authors validated their model performance using 10-fold cross-validation.[37] This approach reduces the effective sample size available for developing the model compared to bootstrap and cross-validation procedures.[36]

External validation is widely recognized as a crucial step in establishing the generalizability and clinical applicability of AI prediction models. External validation is defined as the testing of AI on a new set of data entirely separate from the training dataset, noting it as a crucial step demonstrating that an AI model can work in patients and populations external to the development population.[28] In their scoping review of AI for DR in low-income and middle-income countries, they found that only 43 studies (52%) met their definition of external validation. Prospective validation represents the gold standard for demonstrating real-world clinical utility. For example, a previous study conducted a prospective pilot study with 3,081 patients to demonstrate the

broader applicability of their AI system at the "point-of-care" using fundus images captured with smartphones.[38]

Clinical utility of AI-driven prognostic models

From a clinical and public health perspective, these models can support the early identification of individuals at elevated risk of developing DR before the onset of clinically detectable retinal changes, thereby enabling timely implementation of preventive interventions, optimisation of glycaemic and cardiovascular risk factor control, and prioritisation of ophthalmological surveillance.[39] This is particularly relevant given the growing burden of T2DM and the limited capacity of healthcare systems to provide universal intensive retinal screening. By stratifying patients according to their predicted risk, AI-driven prognostic models may enable more efficient allocation of healthcare resources through risk-based screening strategies, whereby high-risk individuals undergo more frequent monitoring while lower-risk individuals are screened at longer intervals. Such an approach has the potential to improve cost-effectiveness, reduce unnecessary healthcare utilisation, and enhance patient outcomes without compromising safety.[40] Furthermore, the integration of these models within electronic health record systems could facilitate automated risk estimation at the point of care, supporting clinicians in shared decision-making and personalised treatment planning.

Future research direction

Future research should prioritize large-sample, multicentre external validation to address the pervasive high risk of bias identified across existing prediction models. Researchers should develop multimodal models integrating retinal imaging with clinical data, as hybrid approaches consistently enhance predictive performance. Apart from that, future research should also develop dynamic prediction models that can be updated with new patient data over time, enabling continuous risk reassessment and personalized screening intervals. The integration of longitudinal data from electronic health records, continuous glucose monitoring, and serial retinal imaging could enable truly personalized DR management strategies.

LIMITATIONS

Several limitations should be considered when interpreting the findings of this review. First, there was substantial heterogeneity across the included studies in terms of study design, population characteristics, types of AI algorithms, predictor variables, and outcome definitions, which precluded quantitative synthesis and limited direct comparability of results. Second, most studies relied on retrospective or cross-sectional data with predominantly internal validation approaches, while external validation using independent datasets was largely lacking, raising concerns about the generalisability and real-world applicability of the models. Third, reporting of model performance was inconsistent, with a primary focus on discrimination metrics (e.g., AUROC) and limited assessment of calibration and clinical utility. Finally, many studies utilised ML–based feature importance rather than conventional epidemiological effect estimates such as adjusted hazard ratios, restricting interpretability from a clinical perspective.

CONCLUSION

This systematic review demonstrates that AI-driven prognostic models show considerable promise in predicting the risk of DR among individuals with T2DM, with several studies reporting moderate to high predictive performance, particularly for ensemble algorithms such as XGBoost. These models consistently integrate both traditional clinical risk factors and emerging high-dimensional biomarkers, reflecting a shift toward more precise and data-driven risk stratification. However, despite encouraging results, the current evidence is limited by methodological heterogeneity, reliance on internal validation, and insufficient reporting of calibration and clinical utility, which constrain their translation into routine practice. Future research should prioritise standardised reporting, robust external validation, and integration of clinically interpretable models to enhance reliability and facilitate adoption in real-world healthcare settings, ultimately supporting early detection and prevention of DR.

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

FUNDING

The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not- for- profit sectors.

ETHICS APPROVAL

This systematic review followed PRISMA guidelines. It used only publicly available data from the specified databases, with no primary data collection. Ethics approval is not applicable.

REFERENCES

1. Song CM, Lin TH, Huang HT, Yao JY. Illuminating diabetes via multi-omics: Unraveling disease mechanisms and advancing personalized therapy. World J Diabetes 2025; 16(7): 106218.
2. Tahir F, Farhan M. Exploring the progress of artificial intelligence in managing type 2 diabetes mellitus: a comprehensive review of present innovations and anticipated challenges ahead. Front Clin Diabetes Healthc 2023; 4: 1316111.
3. Nimri R, Phillip M, Clements MA, Kovatchev B. Closed-Loop, Artificial Intelligence-Based Decision Support Systems, and Data Science. Diabetes Technol Ther 2025; 27(S1): S64-S78.
4. Dankwa-Mullan I, Rivo M, Sepulveda M, Park Y, Snowdon J, Rhee K. Transforming Diabetes Care Through Artificial Intelligence: The Future Is Here. Popul Health Manag 2019; 22(3): 229-42.
5. Venkatesh R, Gandhi P, Choudhary A, Kathare R, Chhablani J, Prabhu V, et al. Evaluation of Systemic Risk Factors in Patients with Diabetes Mellitus for Detecting Diabetic Retinopathy with Random Forest Classification Model. Diagnostics (Basel) 2024; 14(16).
6. Huang X, Wang H, She C, Feng J, Liu X, Hu X, et al. Artificial intelligence promotes the diagnosis and screening of diabetic retinopathy. Front Endocrinol (Lausanne) 2022; 13: 946915.

7. Yang Q, Bee YM, Lim CC, Sabanayagam C, Yim-Lui Cheung C, Wong TY, et al. Use of artificial intelligence with retinal imaging in screening for diabetes-associated complications: systematic review. EClinicalMedicine 2025; 81: 103089.
8. Ayers AT, Ho CN, Kerr D, Cichosz SL, Mathioudakis N, Wang M, et al. Artificial Intelligence to Diagnose Complications of Diabetes. J Diabetes Sci Technol 2025; 19(1): 246-64.
9. Afrifa-Yamoah E, Adua E, Peprah-Yamoah E, Anto EO, Opoku-Yamoah V, Acheampong E, et al. Pathways to chronic disease detection and prediction: Mapping the potential of machine learning to the pathophysiological processes while navigating ethical challenges. Chronic Dis Transl Med 2025; 11(1): 1-21.
10. Fraser RA, Walker RJ, Campbell JA, Ekwunife O, Egede LE. Integration of artificial intelligence and wearable technology in the management of diabetes and prediabetes. NPJ Digit Med. 2025; 8(1): 687.
11. Corrao S, Janic M, Maggio V, Rizzo M. Machine learning and deep learning in diabetology: revolutionizing diabetes care. Front Clin Diabetes Healthc 2025; 6: 1547689.
12. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021; 372: n71.
13. Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 2021; 11(7): e048008.
14. Yagin FH, Yasar S, Gormez Y, Yagin B, Pinar A, Alkhateeb A, et al. Explainable Artificial Intelligence Paves the Way in Precision Diagnostics and Biomarker Discovery for the Subclass of Diabetic Retinopathy in Type 2 Diabetics. Metabolites 2023; 13(12): 1204.
15. Kotsiliti E, Al-Diri B, Hunter A. A classification model for predicting diabetic retinopathy based on patient characteristics and biochemical measures.pdf. Model Artif Intell Ophthalmol 2017; 1(4): 69-85.
16. Tsao HY, Chan PY, Su EC. Predicting diabetic retinopathy and identifying interpretable biomedical features using machine learning algorithms. BMC Bioinformatics 2018; 19(Suppl 9): 283.
17. Rosu CD, Bratu ML, Stoicescu ER, Iacob R, Hategan OA, Ghenciu LA, et al. Cardiovascular Risk Factors as Independent Predictors of Diabetic Retinopathy in Type II Diabetes Mellitus: The Development of a Predictive Model. Medicina (Kaunas) 2024; 60(10): 1617.
18. Zhao Y, Li X, Li S, Dong M, Yu H, Zhang M, et al. Using Machine Learning Techniques to Develop Risk Prediction Models for the Risk of Incident Diabetic Retinopathy Among Patients With Type 2 Diabetes Mellitus: A Cohort Study. Front Endocrinol (Lausanne) 2022; 13: 876559.
19. Li W, Song Y, Chen K, Ying J, Zheng Z, Qiao S, et al. Predictive model and risk analysis for diabetic retinopathy using machine learning: a retrospective cohort study in China. BMJ Open 2021; 11(11): e050989.
20. Jiang W, Li Z. Comparison of Machine Learning Algorithms and Nomogram Construction for Diabetic Retinopathy Prediction in Type 2 Diabetes Mellitus Patients. Ophthalmic Res 2024; 67(1): 537-48.
21. Kim S, Park J, Son Y, Lee H, Woo S, Lee M, et al. Development and Validation of a Machine Learning Algorithm for Predicting Diabetes Retinopathy in Patients With Type 2 Diabetes: Algorithm Development Study. JMIR Med Inform 2025; 13: e58107.
22. Teruna MMS, Rzazak TR, Yasin SM, Ali AAR, Isa MR. The artificial intelligence driven on the development of diabetic retinopathy prognostic scoring tool among type 2 diabetes mellitus patients: A review. Med J Malaysia 2026; 81: 158-62.
23. Tripp J, Santana-Quinteros D, Perez-Estrada R, Rodriguez-Moran MF, Arcos-Gonzalez C, Mercado-Rios J, et al. Building Machine Learning Models for Type 2 Diabetes Complications. medRxiv. 2023; 10.22.23297277
24. Sheng B, Chen X, Li T, Ma T, Yang Y, Bi L, et al. An overview of artificial intelligence in diabetic retinopathy and other ocular diseases. Front Public Health 2022; 10: 971943.
25. Gong W, Pu Y, Ning T, Zhu Y, Mu G, Li J. Deep learning for enhanced prediction of diabetic retinopathy: a comparative study on the diabetes complications data set. Front Med (Lausanne) 2025; 12: 1591832.
26. Zhang Q, Gong D, Huang M, Zhu Z, Yang W, Ma G. Recent advances and applications of optical coherence tomography angiography in diabetic retinopathy. Front Endocrinol (Lausanne) 2025; 16: 1438739.
27. Vujosevic S, Limoli C, Nucci P. Novel artificial intelligence for diabetic retinopathy and diabetic macular edema: what is new in 2024? Curr Opin Ophthalmol 2024; 35(6): 472-9.
28. Cleland CR, Rwiza J, Evans JR, Gordon I, MacLeod D, Burton MJ, et al. Artificial intelligence for diabetic retinopathy in low-income and middle-income countries: a scoping review. BMJ Open Diabetes Res Care 2023; 11(4): e003424.
29. Huang H, Wu Y, Ye H, Li J, Chen L, Huang X. Risk prediction models for diabetic retinopathy: a systematic review. Front Endocrinol (Lausanne) 2025; 16: 1556049.
30. Peng B, Mu J, Xu F, Guo W, Sun C, Fan W. Artificial intelligence in ophthalmology: a bibliometric analysis of the 5-year trends in literature. Front Med (Lausanne) 2025; 12: 1580583.
31. Bulum T, Tomic M, Vrabec R, Brkljacic N, Ljubic S. Systolic and Diastolic Blood Pressure Are Independent Risk Factors for Diabetic Retinopathy in Patients with Type 2 Diabetes. Biomedicines 2023; 11(8): 2242.
32. Shrateh ON, Abdelhafez M, Ereqat S, Dein LNE, Iriqat S. Identification of Risk Factors for the Development of Diabetic Retinopathy Among Palestinian Adults With Type 2 Diabetes Mellitus: A Cross-Sectional Study. Endocrinol Diabetes Metab 2024; 7(4): e494.
33. Rahman J, Delaforce A, Bradford D, Li J, Magrabi F, Cook D, et al. The Role of Explanations in AI-Generated Alerts: Qualitative Study of Clinical Views on Explainable AI in Predictive Tools. JMIR Hum Factors 2026; 13: e81460.
34. Kanbour S, Harris C, Lalani B, Wolf RM, Fitipaldi H, Gomez MF, et al. Machine Learning Models for Prediction of Diabetic Microvascular Complications. J Diabetes Sci Technol 2024; 18(2): 273-86.
35. Hung KC, Chang LC, Chang YJ, Ho CN, Wu JY, Liu WC, et al. Vitamin D deficiency and diabetic retinopathy risk in patients with newly diagnosed type 2 diabetes mellitus: a retrospective analysis. Front Nutr 2025; 12: 1614287.
36. van Smeden M, Heinze G, Van Calster B, Asselbergs FW, Vardas PE, Bruining N, et al. Critical appraisal of artificial intelligence-based prediction models for cardiovascular disease. Eur Heart J 2022; 43(31): 2921-30.
37. Ke GQ, Fu YJ, Huang ZH, Dai SX, Wen YH, Lv HX. Artificial intelligence in proliferative diabetic retinopathy: advancing diagnosis, precision surgery, and anti-VEGF therapy optimization. Front Med (Lausanne) 2025; 12: 1644456.
38. Bai A, Carty C, Dai S. Performance of deep-learning artificial intelligence algorithms in detecting retinopathy of prematurity: A systematic review. Saudi J Ophthalmol 2022; 36(3): 296-307.
39. Rom Y, Aviv R, Ianchulev T, Dvey-Aharon Z. Predicting the future development of diabetic retinopathy using a deep learning algorithm for the analysis of non-invasive retinal imaging. BMJ Open Ophthalmology 2022; 7(1): e001140.
40. Asif M, Ur Rehman F, Rashid Z, Hussain A, Mirza A, Qureshi WS. An Insight on the Timely Diagnosis of Diabetic Retinopathy Using Traditional and AI-Driven Approaches. IEEE Access 2025; 13: 116869-86.