GMR
The Original

COUNTERFACTUAL LESION REASONING BASED EVIDENCE-VERIFIED RAG FOR EXPLAINABLE DIABETIC RETINOPATHY SEVERITY CLASSIFICATION AND CLINICAL REPORT GENERATION

Dr. N. Balakumar,[1] Vinitha k[2]

[1]Associate Professor and Head, Department of Computer Science , United College of Arts and Science, Coimbatore 20, msg2balakumar@gmail.com
[2]PhD Research scholar, United College of Arts and Science , Coimbatore 20. vini4it@gmail.com

ABSTRACT
Early diagnosis and robust severity classification are significant for providing proper treatment and preventing further damage to vision. Therefore, this research proposes a Counterfactual Lesion Reasoning based Evidence-Verified Retrieval-Augmented Generation (CLR-EVRAG) framework with aiming at enhancing performance of diabetic retinopathy severity classification and clinical report generation. First, retinal fundus images are preprocessed in CLR-EVRAG to find out significant abnormalities, such as microaneurysms, retinal hemorrhages, and hard exudates. The Counterfactual Lesion-Consistency Ordinal Network (CLCON) is then utilized in CLR-EVRAG to classify DR into different severity levels. A counterfactual reasoning mechanism is utilized in CLR-EVRAG to analyze how changes in relevant lesion evidence affect the predicted severity. This process helps assess whether the model's decisions are meaningfully connected to the detected retinal abnormalities and identifies cases that require further investigation. Next, the CLR-EVRAG uses Retrieval-Augmented Generation to obtain relevant information from a carefully selected collection of trusted clinical guidelines and medical references. The retrieval process considers the detected lesions, predicted severity, and evidence required to explain the classification. The retrieved information is then combined with the image-based findings to generate a structured clinical report. An evidence-verification module verifies whether significant statements in the generated report are supported by the retinal image, the retrieved medical sources, or both. Unsupported or uncertain statements are flagged for expert review. The performance of CLR-EVRAG is determined using classification accuracy, retrieval precision, computational complexity, and false positive rate.

KEYWORDS: Diabetic Retinopathy, Severity Classification, Counterfactual Reasoning, Retrieval-Augmented Generation, Evidence Verification, Clinical Report Generation

INTRODUCTION

Diabetic retinopathy (DR) is an eye disease related with diabetes that impacts the blood vessels in the retina. The retina is a light-sensitive layer at the back of the eye that assists a person see. When diabetes impacts these blood vessels, tiny retinal lesions may formulate and gradually lead to vision issues. In advanced cases, the disease can cause severe and permanent vision loss. Early recognition and correct evaluation of disease severity are therefore significant for timely clinical assessment and treatment.

The severity of DR is commonly divided into five levels: no diabetic retinopathy, mild non-proliferative diabetic retinopathy, moderate non-proliferative diabetic retinopathy, severe non-proliferative diabetic retinopathy, and proliferative diabetic retinopathy. These levels represent different stages of retinal damage. Ophthalmologists examine retinal fundus images to identify signs such as microaneurysms, haemorrhages, hard exudates, and other abnormal changes. However, accurate grading can be difficult because the appearance, number, size, and distribution of these lesions vary between patients. Differences in image quality, illumination, contrast, and retinal visibility can also affect the reliability of the assessment.

In addition to accurate classification, generating a useful clinical report is an important research goal. A report can summarize the image quality, detected retinal abnormalities, predicted severity, and relevant limitations of the automated assessment. Retrieval-augmented generation (RAG) can support report preparation by retrieving information from trusted clinical guidelines and reviewed medical publications before generating text. However, retrieved information does not automatically make a report correct. A language model may include a statement that is unsupported by the retinal image, misinterpret a source, or produce a description that conflicts with the predicted grade. Therefore, generated reports require a verification process that checks whether image-related claims agree with model outputs and whether clinical statements are supported by appropriate references.

In the state-of-the-art works, Alshehri and Basbrain [1] implemented a CNN-RAG method for DR severity assessment and clinical report generation (CRG) by combining convolutional neural network (CNN) classification with retrieval-augmented generation (RAG). This approach connects image-based severity prediction with automatically generated medical descriptions, but the extent to which it verifies each report statement against image findings and retrieved evidence needs to be examined. Zaian et al. [2] introduced Retina-RAG, which combines a retinal image classifier; a vision-language model fine-tuned using Low-Rank Adaptation (LoRA), and a knowledge retrieval module to generate clinical

reports. Its modular design supports disease grading and report generation, although the need for stronger counterfactual analysis and explicit claim-level evidence verification remains an important area to investigate.

In order to resolve these challenges, this CLR-EVRAG proposes a Counterfactual Lesion-Consistency Ordinal Network (CLCON) as the classification component of an explainable DR assessment and report-generation framework. The proposed network is designed to combine vital retinal features with lesion-related features and predict severity using an ordinal learning objective. A counterfactual consistency component is introduced to evaluate whether the model responds appropriately to carefully design and clinically reviewed changes in lesion-related information. The classification output, detected lesion evidence, and relevant uncertainty information are then used by the wider framework to retrieve supporting clinical evidence and generate a report. Before the report is presented, verification stage checks consistency between the predicted grade, image-derived findings, and retrieved clinical information.

1.1 Research Gap

References [1] and [2] explore automated DR severity assessment and retrieval-augmented CRG. However, further research is needed to ensure that predicted severity levels are strongly supported by detected retinal lesions and reliable clinical evidence. The integration of counterfactual lesion reasoning, ordinal severity classification, and claim-level verification of generated reports remains an important area for investigation. Therefore, the proposed work aims to address these challenges by developing a lesion-aware classification approach and an evidence-verified RAG framework to improve prediction consistency, explainability, and the reliability of clinical reports.

1.2 Contributions of Research Work

The main contributions of this research are as follows:

- A novel CLCON classification approach is designed to combine key retinal features and lesion-related evidence while accounting for the ordered relationship between the five DR severity levels.
- Counterfactual lesion-consistency analysis is introduced to examine how controlled, clinically reviewed changes in lesion-related features influence severity predictions. This component aims to improve the consistency and interpretability of model behavior.
- A verification stage is proposed to identify unsupported clinical statements, inconsistencies between the predicted grade and image findings, and claims that lack adequate source evidence.

2. LITERATURE REVIEW

Ito et al. [3] developed XDR-LVLM, an explainable vision-language model that uses a medical vision encoder, multi-task prompt engineering, and multi-stage fine-tuning to classify DR severity and describe related retinal abnormalities. It provides understandable explanations linked to detected pathological concepts, but the authors identify difficulties in detecting subtle lesions and the need for further uncertainty assessment and clinical validation Agarwal et al. [4] proposed a lesion-aware ordinal transformer that combines morphological lesion extraction, ConvNeXt-V2, a Swin-V2 transformer, and ordinal learning for five-level DR grading. The method focuses on small retinal lesions and respects the ordered nature of disease severity, but its primary focus is classification and visual explanation rather than counterfactual reasoning and evidence-verified CRG.

Akhtar et al. [5] developed RSG-Net, a convolutional neural network using histogram equalization, image de-noising, data augmentation, and convolutional feature extraction for DR grading on the Messidor-1 dataset. The model reported high classification accuracy, but its main focus is severity prediction; lesion-level counterfactual explanations and evidence-supported narrative reporting are not considered. Ikram and Imran [6] presented ResViT FusionNet, a hybrid model that combines ResNet50 and Vision Transformer (ViT) for DR severity grading. The method uses data augmentation and explainable AI techniques, including LIME and Grad-CAM, to highlight image regions influencing predictions. However, these visual explanations do not by themselves verify whether every clinical statement is supported by the detected lesions or medical evidence.

Alamri et al. [7] developed an ensemble framework combining Graph Attention Networks (GATs), Graph Convolutional Networks (GCNs), and GraphSAGE, with a Transformer encoder for feature fusion. The method also uses a conditional generative adversarial network to address class imbalance and reported promising classification results. However, the use of several neural network models and synthetic images may increase computational demands. Aftab and Akhtar [8] designed an ensemble transfer-learning method using EfficientNetB2, DenseNet121, and ResNet50 for five-level DR grading. Their approach combines fundus image datasets, applies contrast enhancement and data augmentation, and averages the predictions of the three models to improve classification performance. However, differences in image quality and annotation standards across datasets may affect consistency.

Saleem et al. [9] introduced CG-DRNet, a clinical-guided deep learning framework that combines lesion-aware attention, adversarial data augmentation, and uncertainty estimation using Monte Carlo dropout. It also uses a multi-task learning strategy to identify retinal lesions and an uncertainty-informed learning approach to improve prediction reliability. Although these features support interpretable and uncertainty-aware DR grading, the framework does not primarily focus on counterfactual lesion reasoning. Rezaee and Farnami [10] intended a hybrid CNN–Transformer model that combines local image features extracted by convolutional neural networks with broader contextual information learned by Transformer layers. However, lesion-based counterfactual explanations and verification of clinical statements against retrieved medical evidence not focused.

Yan et al. [11] conducted a systematic review and meta-analysis to examine the diagnostic accuracy and clinical performance of deep learning models for DR grading. However, differences in datasets, grading criteria, and evaluation methods can make direct comparisons difficult, while the review does not itself provide a new lesion-based counterfactual reporting framework. Swapna et al. [12] implemented a DR detection approach that combines federated learning with Vision Transformers. Federated learning allows models to be trained across different data sources without directly sharing

the original retinal images, while Vision Transformers learn important image features for classification. Though, lack of explicit evidence verification in CRG remains open issue.

Yadav et al. [13] presented a deep learning approach for classifying DR from retinal fundus images using transfer learning. The method reuses features learned by a pre-trained model and adapts them to the DR classification task, reducing the need to train a complete network from the beginning. However, the classification output alone may not fully explain which lesions influenced the predicted grade, creating scope for lesion-level reasoning and evidence-supported clinical explanations. Vohra et al. [14] developed a low-cost AI-based system for early detection of DR and other ocular diseases, with a focus on resource-limited healthcare settings. The approach aims to make automated eye screening more accessible by considering practical deployment requirements and the detection of multiple eye conditions. However, the reliability of its outputs across different clinical environments was considered.

Zhang et al. [15] introduced IRG-DR, an interactive report generation framework that uses evidence guidance to support DR-related clinical reporting. Its emphasis on connecting report generation with supporting evidence is relevant to improving the reliability of automated medical descriptions. However, the extent to which counterfactual lesion changes are used to examine severity predictions was remained unsolved. Huang [16] developed DeepEyeNet, a framework for generating medical reports from retinal images by integrating visual information from retinal images with language generation to create structured descriptions, helping minimize the requirement of fully manual report construction. However, automated descriptions may necessitate additional checks to make sure that the reported findings match the image and that clinically vital statements are supported by reliable evidence.

Palaniappan et al. [17] investigated the use of generative AI to produce realistic synthetic images for DR analysis. Synthetic image generation can increase the diversity of training data and may help address limited representation of certain disease grades. However, generated images may contain unrealistic retinal structures or inaccurate lesion patterns, so their clinical quality must be checked before they are used for training or evaluation. Gundu et al. [18] studied the efficiency of RAG for DR categorization. Their work explores whether adding retrieved information can support the classification process and improve the use of relevant clinical knowledge. However, the benefit of retrieval depends on the quality and relevance of the retrieved information.

Zhou et al. [19] presented a comparative evaluation of multi-agent systems for diabetic retinal disease detection and CRG. This comparison helps investigate whether dedicated visual analysis tools improve the interpretation of retinal images and the quality of generated reports. However, the reliability of individual report statements and the contribution of counterfactual lesion reasoning require separate assessment beyond overall system performance. Han et al. [20] studied the drawbacks of RAG-based medical vision-language models through a case study in ophthalmic report generation. Their work is significant to understanding how retrieved information interacts with visual evidence and how retrieval-based systems may create incomplete or unsupported explanations. This motivates the use of explicit verification procedures that compare generated claims with retinal findings and source documents before a report is finalized. The following table shows the comparative review of existing research on DR detection and CRG.

Table 1 Comparative Literature Table

Ref.	Author(s) and Year	Technique Used	Main Advantage	Limitation / Research Gap
[1]	Alshehri and Basbrain (2026)	CNN-based DR grading combined with RAG for narrative report generation.	Combines severity prediction with automated clinical descriptions.	The extent of lesion-level reasoning and claim-level evidence verification needs further investigation.
[2]	Zaian et al. (2026)	Retina-RAG, a retrieval-augmented vision-language framework for retinal diagnosis and report generation.	Connects retinal image analysis with retrieved ophthalmic knowledge.	Counterfactual lesion analysis and explicit verification of individual report claims require further study.
[3]	Ito et al. (2025)	XDR-LVLM, an explainable vision-language model for DR diagnosis.	Connects disease prediction with descriptions of retinal abnormalities.	Further evaluation of subtle lesion detection, uncertainty, and evidence-grounded reporting may be needed.
[4]	Agarwal et al. (2026)	Lesion-aware ordinal Transformer for fundus-based DR grading.	Considers lesion features and the ordered nature of DR severity grades.	Counterfactual explanations and verified CRG are not the primary focus.
[5]	Akhtar et al. (2025)	Deep learning model for DR grading using fundus images and image-processing techniques.	Supports automated severity classification from retinal images.	Classification results may not provide sufficient lesion-level explanations or verified narrative reports.
[6]	Ikram and Imran (2025)	ResViT FusionNet combining CNN and Vision Transformer features with explainability methods.	Combines local and global image features and provides visual explanations.	Highlighted regions do not necessarily establish that every clinical statement is supported by evidence.

[7]	Alamri et al. (2026)	Ensemble of graph neural networks with Transformer-based feature fusion.	Combines different feature-learning methods for DR severity classification.	Model complexity and the absence of integrated counterfactual report verification require consideration.
[8]	Aftab and Akhtar (2025)	Data fusion and ensemble transfer learning for DR classification.	Combines pretrained models and information from multiple data sources.	Differences between datasets may affect generalization; evidence-grounded report generation is not the main objective.
[9]	Saleem et al. (2026)	Clinical-guided learning with lesion-aware attention, adversarial augmentation, and uncertainty quantification.	Combines lesion information and uncertainty estimation for DR assessment.	The integration of counterfactual lesion reasoning with claim-level verification in generated reports needs investigation.
[10]	Rezaee and Farnami (2025)	CNN–Transformer fusion for DR severity classification.	Combines detailed local features with broader image context.	Classification-focused outputs may not explain the effect of individual lesions or provide verified clinical reports.
[11]	Yan et al. (2026)	Systematic review and meta-analysis of deep learning models for DR grading.	Summarizes diagnostic accuracy and clinical performance across existing studies.	Differences in datasets and evaluation methods may limit direct comparisons; it does not introduce a new diagnostic framework.
[12]	Swapna et al. (2025)	Federated learning combined with Vision Transformers for DR detection.	Supports collaborative model training without directly sharing original retinal images.	Differences in data distributions and the lack of integrated report verification remain relevant challenges.
[13]	Yadav et al. (2025)	Transfer-learning-based DR classification using fundus images.	Reuses pre-trained features to support efficient model development.	May provide limited lesion-level explanations and does not primarily address evidence-verified reporting.
[14]	Vohra et al. (2025)	Low-cost AI-based system for early DR and ocular disease detection.	Focuses on accessible screening in resource-limited settings.	Generalization across clinical environments and integration with counterfactual, evidence-supported reporting need further evaluation.
[15]	Zhang et al. (2026)	IRG-DR, an interactive report-generation framework with evidence guidance.	Incorporates evidence guidance into DR clinical report generation.	Its relationship to counterfactual lesion reasoning and claim-level verification against both image findings and retrieved sources requires detailed comparison.
[16]	Huang (2025)	DeepEyeNet for medical report generation from retinal images.	Supports automated generation of retinal image descriptions.	Generated findings may require additional image-grounding and clinical evidence checks.
[17]	Palaniappan et al. (2025)	Generative AI for realistic synthetic retinal image generation.	May increase training-data diversity and support underrepresented classes.	Synthetic images may contain unrealistic or misleading retinal features and require clinical quality assessment.
[18]	Gundu et al. (2026)	Evaluation of RAG for DR classification.	Investigates the value of retrieved knowledge for DR classification.	Retrieval quality and the consistency between retrieved evidence and image-specific findings require careful evaluation.
[19]	Zhou et al. (2026)	Comparative evaluation of multi-agent systems with and without specialized image-based tools.	Examines the role of visual tools in retinal disease detection and report generation.	The accuracy of individual report claims and the contribution of counterfactual explanations need separate evaluation.
[20]	Han et al. (2025)	Analysis of constraints in RAG-based medical vision-language models for ophthalmic reporting.	Highlights challenges in retrieval-based medical report generation.	Supports the need for explicit checks of generated claims against visual findings and retrieved source evidence.

3. PROPOSED METHODOLOGY

This paper implements a novel CLR-EVRAG framework for DR severity classification and CRG using counterfactual lesion reasoning and evidence-verified retrieval-augmented generation. The key objective of CLR-EVRAG is to classify DR severity, detect significant retinal lesions, explain the model's predictions, and make reports supported by clinical

evidence. The vital contribution of the CLR-EVRAG is to use counterfactual lesion analysis to connect DR severity classification with clinical evidence retrieval. After predicting retinal abnormalities and disease grade, the CLR-EVRAG determines how controlled changes to chosen lesion features influence the model's result. This process assists CLR-EVRAG to examine whether the prediction is reliable with the identified retinal findings. The information obtained from this analysis is then utilized to retrieve clinical references that are associated to the discovered grade and observed lesions. From that, evidence retrieval is guided by both the classification output and the lesion-based explanation.

Another contribution is the verification of individual claims before generating the final clinical report. Although a language model may generate fluent medical text, some statements may not be adequately supported by the retinal image or the retrieved clinical information. To deal with this problem, the proposed CLR-EVRAG verifies whether reported lesions agree with the image analysis, whether the stated DR grade matches the classifier's prediction, and whether clinical explanations are supported by the retrieved references. Unsupported clinical statements are revised or flagged for further examination. Thus, proposed CLR-EVRAG improves accuracy, consistency, and traceability of automatically generated clinical reportsThe architecture of CLR-EVRAG framework is shown in Figure 1.

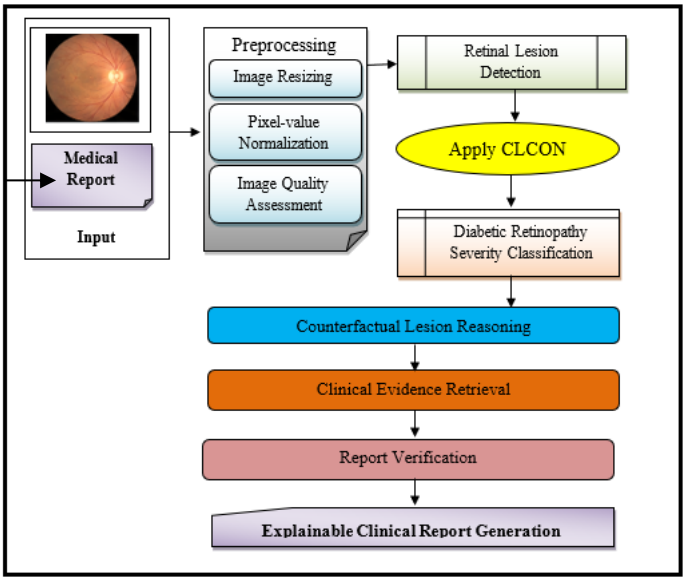


Figure 1 Workflow of CLR-EVRAG architecture

As presented in above Figure 1, CLR-EVRAG framework includes of seven main stages i.e. retinal fundus image preprocessing, lesion detection, DR severity classification, counterfactual lesion reasoning, clinical evidence retrieval, report verification, and explainable CRG. Each stage provides information to the next stage to support a more transparent and reliable assessment. The complete workflows of CLR-EVRAG framework are described as follows.

3.2 Retinal Fundus Image Preprocessing

The first stage of CLR-EVRAG receives a retinal fundus image as input and preprocesses it for further analysis. Image quality is a vital factor in DR recognition because poor illumination, blur, low contrast, and incomplete retinal views may affect lesion discovery and severity categorization. For that reason, preprocessing and image quality assessment are carried out before the image is passed to the next stage. Let the input retinal fundus image be described by I where $I \in R^{H\times W\times C}$. In CLR-EVRAG, H point outs the image height, W defines the image width, and C refers the number of colour channels. The preprocessing operation is mathematically performed as,

$$I_P = \mathcal{P}(I) \quad (1)$$

In equation (1), $\mathcal{P}(.)$ indicates the preprocessing function and I_P describes the preprocessed retinal fundus image. During this task, image resizing is mathematically done with the support of below equation,

$$I_r = Resize(I, H_t, W_t) \quad (2)$$

In equation (2), I_r depicts the resized image, and H_t and W_t indicate the target height and width, respectively. Then, pixel-value normalization is mathematically performed as,

$$x_n = \frac{x - x_{min}}{x_{max} - x_{min+\varepsilon}} \quad (3)$$

In equation (3), x is the original pixel value, x_{min} and $x_{max}$are the minimum and maximum pixel values, x_n is the normalized pixel value and $\varepsilon > 0$ is a small constant that prevents division by zero. Followed by, image quality assessment is mathematically done as,

$$Q(I_p) = f_p(I_p) \quad (4)$$

In equation (4), $f_p(.)$denotes the image quality assessment function and $Q(I_p)$ represents the quality score of the preprocessed image. After that, quality-based decision is mathematically performed as,

$$D_q = \begin{cases} 1, & If Q(I_p) \geq \tau_q \\ 0, & If Q(I_p) \geq \tau_q \end{cases} \quad (5)$$

In equation (5), $D_q = 1$ indicates acceptable image quality, $D_q = 0$ indicates that the image requires further review, and τ_q is the quality threshold determined using validation data. Finally, the quality images that are selected to the subsequent lesion detection stage is mathematically obtained as,

$$I_{out} = \begin{cases} I_p, & D_q = 1 \\ Refer\ for\ Quality\ Review, & D_q = 0 \end{cases} \quad (6)$$

In equation (6), I_{out}represents the output of the preprocessing and quality assessment stage. Images that meet the quality requirement are passed to the subsequent lesion detection stage, while images that fail the quality check are flagged for review. Thus, the preprocessing and quality assessment stages aim to provide consistent retinal images for lesion recognition and DR severity grading.

3.2 Retinal Lesion Detection

The second stage of the proposed framework identifies retinal lesions associated with DR from the preprocessed fundus image. These lesions provide important information for understanding disease severity and explaining the classification results. The target lesions may include microaneurysms, haemorrhages, hard exudates, and other retinal abnormalities, depending on the selected dataset and the availability of lesion annotations. Let the preprocessed retinal image be represented byI_p. A lesion detection model is applied to this image to identify the locations and types of retinal abnormalities. The lesion detection process is mathematically expressed as,

$$L = f_\theta(I_p) \quad (7)$$

In equation (7), $f_\theta(.)$denotes the trained lesion detection model, θ represents its learnable parameters, and L denotes the set of detected retinal lesions. The output of the retinal lesion detection model is mathematically represented as,

$$L = \{(c_i, b_i, s_i)\}_{i=1}^{N} \quad (8)$$

In equation (8), N is the number of detected lesions, c_i indicates the predicted lesion class, b_i represents the bounding box or spatial location of the i^{th} lesion, and s_i represents its confidence score. If a segmentation model is used instead of a bounding-box detector, b_i can be replaced by a binary lesion mask M_i, which discovers pixels associated with the lesion. For segmentation-based lesion recognition, the predicted mask for lesion class (k) is mathematically formulated as,

$$M_k(x,y) = \begin{cases} 1, If\ P_k(x,y) \geq \tau_k \\ 0, If\ P_k(x,y) < \tau_k \end{cases} \quad (9)$$

In equation (9), $P_k(x,y)$ describes predicted probability that pixel (x,y)belongs to lesion class k and τ_k defines the decision threshold. A value of 1 represents that the pixel is assigned to the lesion class, while a value of 0 point outs that it is not assigned to that class. The discovered lesions are then considered as a lesion feature vector for DR severity classification and counterfactual reasoning stages. From that, lesion feature extraction is mathematically represented as,

$$z_L = \varphi(L) \quad (10)$$

In equation (10), $\varphi(L)$ depicts the lesion feature extraction and z_L illustrates the resulting lesion-related features. These features may comprise lesion type, location, estimated area, distribution, and detection confidence, depending on the model result. For segmentation, the Dice coefficient is utilized to determine the overlap between predicted lesion mask and reference mask which mathematically obtained as,

$$Dice(M_p, M_g) = \frac{2|M_p \cap M_g|}{|M_p| + |M_g| + \varepsilon} \quad (11)$$

In equation (11), M_p is the predicted lesion mask, $M_g$is the ground-truth mask, $|.|$denotes the number of pixels in a mask, and ε is a small positive constant used to prevent division by zero. The Dice coefficient result ranges from 0 to 1 where higher values indicating greater overlap between the predicted and reference masks. This stage provides structured information about the retinal lesions visible in the input image. These findings help link the predicted DR severity to image-based evidence and provide the lesion information required for counterfactual reasoning and evidence-supported clinical report generation.

3.3 Diabetic Retinopathy Severity Classification

After lesion analysis, the CLCON is used to detect the severity level of DR. The classification model receives the retinal image and may also use lesion-related features, such as lesion type, location, distribution, and estimated quantity. DR severity classification is an ordinal learning problem because the five severity grades follow a natural order. The grades are defined as 0 (no DR), 1 (mild non-proliferative DR), 2 (moderate non-proliferative DR), 3 (severe non-proliferative DR), and 4 (proliferative DR). The CLCON used in this proposed work estimates four cumulative probabilities which mathematically obtained as,

$$q_k(I_p) = P(Y > k | I_p) = \sigma(w^T F_{fused} - b_k), k = 0,1,2,3 \quad (12)$$

In equation (12), Y describes true severity grade, w refers a learnable weight vector, b_k represents the ordered thresholds, and $\sigma(.)$ point outs sigmoid function. The thresholds satisfies the below following mathematical equation,

$$b_0 \leq b_1 \leq b_2 \leq b_3 \quad (13)$$

This constraint ensures that the cumulative probabilities follow the expected order given below,

$$q_0 \leq q_1 \leq q_2 \leq q_3 \quad (14)$$

The probability of each individual severity grade is then mathematically determined as,

$$P(Y = r | I_p) = \begin{cases} 1 - q_{0,} & r = 0 \\ q_{r-1} - q_{r,} & r \in \{1,2,3\} \\ q_{3,} & r = 4 \end{cases} \quad (15)$$

The predicted severity grade is mathematically measured from the class with the highest estimated probability which mathematically expressed as,

$$\hat{Y} = arg \max_{r \in \{0,1,2,3\}} P(Y = r | I_p) \quad (16)$$

This ordinal formulation accounts for the ordered relationship between severity grades and provides a probability distribution over the five classes. These outputs are passed to the counterfactual reasoning stage.

3.4 Counterfactual Lesion Reasoning

In the fourth stage, counterfactual reasoning is introduced in CLR-EVRAG to investigate how changes in selected lesion-related information influence the model's severity prediction. A counterfactual input is generated by applying a controlled transformation to the original image or its lesion representation which mathematically described as,

$$I^{cf} = T(I_p, L, \delta) \quad (17)$$

In equation (17), I^{cf} depicts counterfactual input generated for analyzing changes in the predicted DR severity, $T(.)$ point outs counterfactual transformation function utilized to modify chosen lesion-related information, I_p refers original preprocessed retinal fundus image, and L indicates set of detected retinal lesions, such as microaneurysms, haemorrhages, and hard exudates. Here, δ defines controlled changes applied to selected lesion-related information, such as modifying lesion presence, size, or severity-related features. The transformation should protect relevant retinal anatomy and avoid introducing artificial patterns that could mislead the classifier. Where realistic image-level counterfactuals cannot be generated reliably, controlled changes to lesion features may be used instead. Each transformation should be reviewed or validated using appropriate clinical criteria. A counterfactual experiment represents a controlled model analysis and does not, by itself, establish a causal relationship between a lesion and disease severity.

3.5 Clinical Evidence Retrieval Using RAG

The fifth stage uses RAG to retrieve relevant clinical information from a trusted knowledge base containing DR grading guidelines, ophthalmology references, and reviewed clinical publications. The retrieval query is constructed using the predicted severity grade, detected lesions, counterfactual findings, and model uncertainty. Relevant evidence is selected to support the clinical report, and the source details are recorded for verification. The retrieval process is mathematically expressed as,

$$E_r = TopK(Sim(Q, K_i)) \quad (18)$$

In equation (18), E_r demonstrates retrieved clinical evidence passages, Q refers retrieval query generated from the model's findings, K_i and represents individual documents or passages in the clinical knowledge base, $Sim(Q, K_i)$ describes similarity score between the query and each passage whereas $TopK(.)$ used to selects the k most relevant passages. The retrieved evidence presents clinical context for explaining the detected DR severity.

3.6 Clinical Report Verification

Before constructing the final clinical report, the CLR-EVRAG framework verifies whether the proposed statements are supported by the detected retinal lesions, predicted DR severity, counterfactual findings, and retrieved clinical evidence. It checks lesion precision, severity agreement, evidence support, consistency, and uncertainty. Unsupported or contradictory statements are revised or removed, while uncertain findings are considered for further clinical review. The report verification process is mathematically performed as,

$$V_r = F_v \wedge S_v \wedge E_v \wedge C_v \wedge U_v \quad (19)$$

In equation (19), V_r includes overall report verification status, F_v describes lesion finding verification, S_v represents DR severity verification, E_v defines clinical evidence verification, clinical evidence verification, C_v point outs consistency verification and U_v indicates uncertainty assessment. Here, $\wedge$ demonstrates the logical AND operation. The report is considered as fully verified only when all required verification conditions are fulfilled. If any condition fails, the relevant statement is considered for correction or further review.

The proposed verification procedure includes the following checks:

- **Finding Verification**: Confirm that statements about retinal lesions are supported by the lesion detection results.
- **Severity Verification**: Check that the reported DR grade matches the classification output and the grading criteria used.
- **Evidence Verification**: Confirm that clinical explanations are supported by the retrieved source passages.
- **Consistency Verification**: Identify contradictions between lesion findings, the predicted grade, and the proposed report.
- **Uncertainty Verification**: Identify statements that cannot be supported confidently because of poor image quality, uncertain predictions, or insufficient evidence.

3.8 Explainable Clinical Report Generation

In the final stage, the verified information is considered to prepare an explainable clinical report. The report is designed to present the findings in clear language and to differentiate image-based observations from information acquired from clinical references. The report may contain the following components:

- **Image Quality**: A summary of whether the image is suitable for automated assessment.
- **Detected Lesions**: The lesion types and locations identified by the model, with visual evidence where available.
- **Predicted DR Severity**: The predicted grade and associated confidence information.
- **Counterfactual Explanation**: A summary of how controlled changes to selected lesion evidence affected the model's prediction.
- **Clinical Evidence**: Relevant grading information retrieved from trusted sources, including source references.
- **Verification Status**: An indication of whether the main report claims are supported, uncertain, or unresolved.
- **Recommended Next Step**: A cautious suggestion for further assessment or specialist review when the image, prediction, or evidence is insufficient.

The report should avoid presenting model predictions as confirmed clinical diagnoses. Cases with poor image quality, low confidence, conflicting evidence, or unresolved verification results should be flagged for assessment by a qualified eye-

care professional. The CLR-EVRAG framework is intended to support clinical assessment rather than replace professional judgment. The algorithmic processes CLR-EVRAG framework is presented as follows,

Input: Retinal fundus image, Patient medical report
Output: DR severity class, Detected retinal lesions, Explainable clinical report
Begin
Step 1: For each input retinal fundus image and medical report
Step 2: Perform preprocessing using (1) to (3)
Step 3: Perform image quality assessment using (4) to (6)
Step 4: Lesion detection model extracts important retinal regions using (7) to (11)
Step 5: Apply CLCON
Step 6: Predicts the disease stage using (12) to (16)
Step 7: Cunterfactual reasoning examine how changes in selected lesion-related information influence the model's severity prediction using (17)
Step 8: Predicted severity grade, detected lesions, counterfactual findings, and model uncertainty are combined to construct the retrieval query
Step 9: Retrieve relevant clinical information using (18)
Step 10: Verifies whether the proposed statements are supported by detected retinal lesions, predicted DR severity, counterfactual findings, and retrieved clinical evidence using (19)
Step 11: Generates an explainable clinical report
Step 12: End For
End

Algorithm 1 Lesion Aware Uncertainty Adaptive Retrieval-Augmented Generation Algorithm

4. Experimetnal Setup

The CLR-EVRAG framework is implemented in Python by considering FFA-IR datasets input. This dataset comprises 10,790 clinical reports and 1,048,584 fundus images gathered from real clinical practice The experimental process includes retinal image quality assessment, lesion detection, DR severity classification, counterfactual lesion reasoning, retrieval-augmented clinical evidence generation, and report verification. The dataset is partitioned into training, validation, and testing subsets at the patient or case level to reduce data leakage. The performance is evaluated using classification accuracy, retrieval precision, computational complexity, and false positive rate. The generated reports are further examined for factual consistency, clinical evidence support, and unsupported statements. The experimental results of proposed CLR-EVRAG are compared against with conventional CNN-RAG [1] and Retina-RAG [2].

4.1 Test Case 1: Classification Accuracy

Classification accuracy measures the proportion of correctly predicted samples out of the total number of samples evaluated. In the proposed CLR-EVRAG framework, it indicates how accurately the model predicts the DR severity grades compared with the actual expert-assigned labels. A higher accuracy indicates that the model correctly classifies a larger proportion of the test images.

Table 1 Simulation Performance of Accuracy for DR Classification

Number Of Fundus Images	Classification Accuracy (%)		
	CNN-RAG	Retina-RAG	Proposed CLR-EVRAG
2000	93.05	95.45	97.06
4000	93.42	95.74	97.48
6000	93.74	96.08	97.93
8000	93.91	96.62	98.12
10000	94.23	97.27	98.64

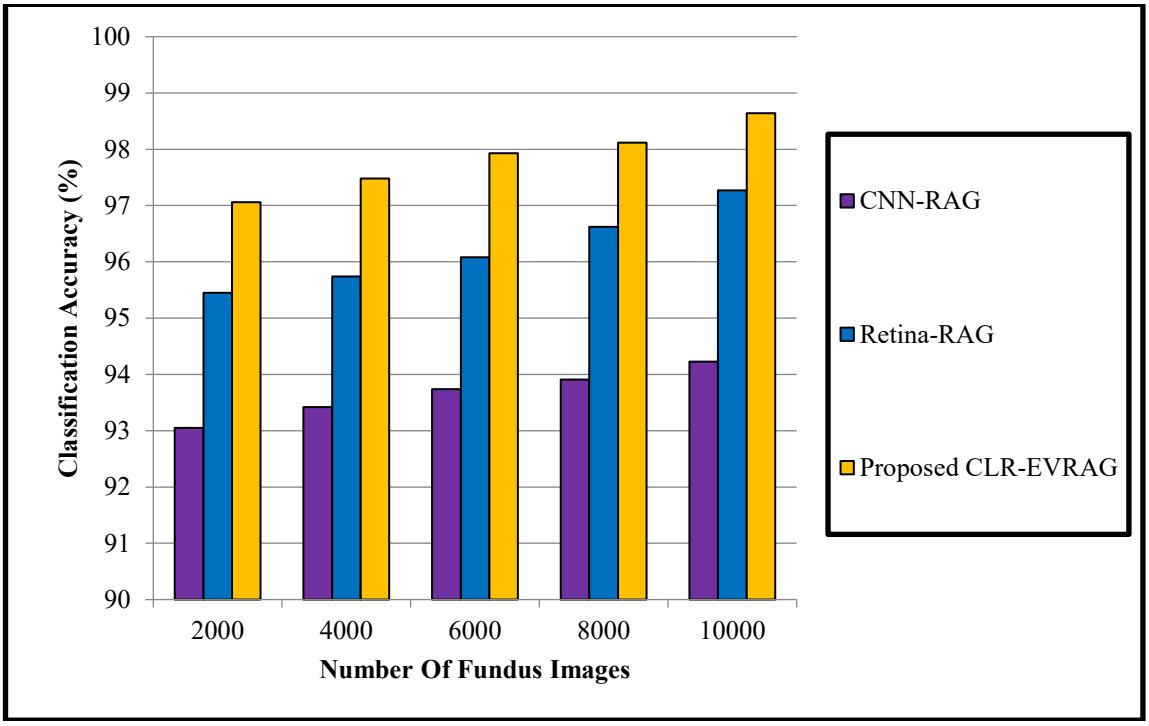


Figure 2 Graphical Representation of Classification Accuracy

Table 1 and Figure 2 compares the classification accuracy of three methods: CNN-RAG, Retina-RAG, and the proposed CLR-EVRAG framework using different numbers of fundus images, ranging from 2,000 to 10,000. The classification accuracy is measured in percentage, where a higher value indicates better prediction performance. The conventional CNN-RAG [1] achieves an accuracy of approximately 93.1% to 93.5% as the number of images increases whereas existing Retina-RAG [2] performs better, with accuracy improving from approximately 95.4% to 96.7%. However, the proposed CLR-EVRAG framework achieves the highest accuracy among all three methods, increasing from approximately 97.0% with 2,000 images to around 98.1% with 10,000 images. The graph shows that classification accuracy improves as the number of fundus images increases for all three methods. The proposed CLR-EVRAG framework maintains the best performance across all five dataset sizes. Its higher accuracy may be associated with the combined use of lesion-based analysis, counterfactual reasoning, clinical evidence retrieval, and report verification. These components are designed to support more reliable diabetic retinopathy severity prediction. Overall, the results shown in the graph suggest that CLR-EVRAG performs better than CNN-RAG and Retina-RAG in classification accuracy.

4.2 Test Case 2: Retrieval Precision

Retrieval precision measures the proportion of retrieved clinical evidence passages that are relevant to the given query. In the proposed CLR-EVRAG framework, it evaluates how effectively the RAG component retrieves useful clinical information related to the predicted DR severity, detected retinal lesions, and clinical grading criteria. A higher retrieval precision indicates that a larger proportion of the retrieved passages are relevant to the clinical query, helping the framework generate reports supported by appropriate clinical evidence.

Table 2 Simulation Performance of Retrieval Precision for Clinical Report Generation

Number Of Fundus Images	Retrieval Precision (%)		
	CNN-RAG	Retina-RAG	Proposed CLR-EVRAG
2000	92.14	93.43	95.63
4000	92.58	93.88	95.97
6000	92.97	94.08	96.49
8000	93.46	94.64	96.95
10000	94.02	95.57	98.02

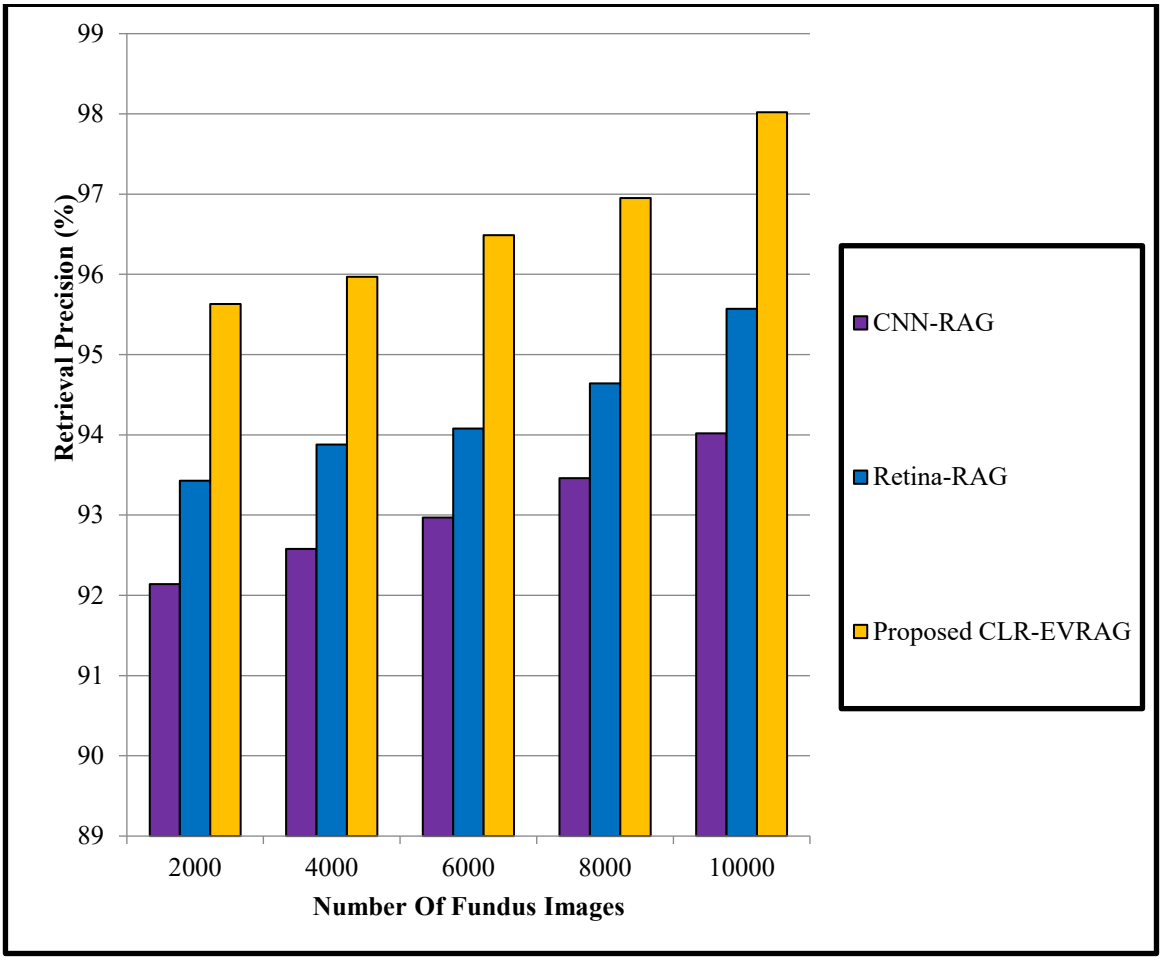


Figure 3 Graphical Representation of Retrieval Precision

Table 2 and Figure 3 demonstrates the retrieval precision of three methods, namely CNN-RAG, Retina-RAG, and the proposed CLR-EVRAG framework, using different numbers of fundus images ranging from 2,000 to 10,000. Retrieval precision measures how many of the retrieved clinical evidence passages are relevant to the given query. A higher percentage indicates that the retrieval system selects more relevant information. The state-of-the-art CNN-RAG [1] achieves a retrieval precision of approximately 92.2% to 93.3% as the number of images increases whereas conventional Retina-RAG [2] shows better performance, improving from approximately 93.4% to 94.9%. The proposed CLR-EVRAG framework achieves the highest retrieval precision, increasing from approximately 95.5% for 2,000 images to around 97.5% for 10,000 images. The graph indicates that retrieval precision improves with increasing dataset size for all three methods. CLR-EVRAG maintains the highest values across all five image counts. This improvement may be related to its use of predicted DR severity, detected lesion information, counterfactual findings, and uncertainty information to guide clinical evidence retrieval. These features are intended to help the framework identify clinical references that are more closely related to the observed retinal findings. Overall, the graph suggests that CLR-EVRAG retrieves relevant clinical evidence more effectively than CNN-RAG and Retina-RAG.

4.3 Test Case 3: Computational Complexity

Computational complexity refers to the total time required by the proposed CLR-EVRAG framework to process a retinal image, detect lesions, predict diabetic retinopathy (DR) severity, perform counterfactual reasoning, retrieve relevant clinical evidence, and verify the generated report. It is used to assess the processing efficiency of the framework.

Table 3 Simulation Performance of Computational Time

Number Of Fundus Images	Computational Time (ms)		
	CNN-RAG	Retina-RAG	Proposed CLR-EVRAG
2000	22.9	20.2	16.1
4000	24.8	22.8	18.7
6000	29.7	25.6	21.2
8000	32.5	30.7	26.3
10000	36.4	34.9	29.5

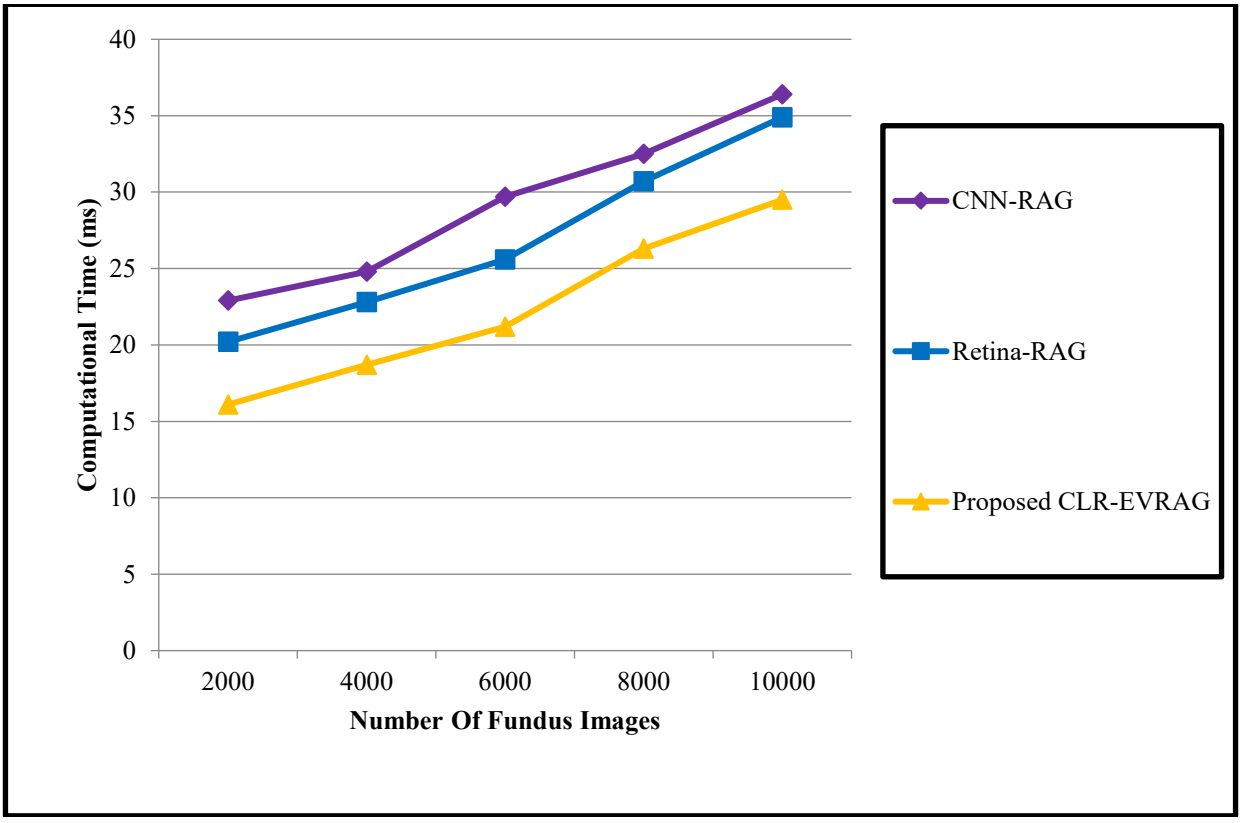


Figure 4 Graphical Representation of Computational Time

Table 3 and Figure 4 depict computational time of three methods: CNN-RAG, Retina-RAG, and the proposed CLR-EVRAG for different numbers of fundus images. The number of images ranges from 2,000 to 10,000, and the computational time is measured in milliseconds (ms). As shown in the graph, the computational time increases as the number of fundus images increases for all three methods. The existing CNN-RAG [1] takes the highest time, increasing from approximately 22.8 ms for 2,000 images to 36.5 ms for 10,000 images. Retina-RAG [2] shows a lower computational time than CNN-RAG [1], rising from about 20.2 ms to 35 ms. In comparison, the proposed CLR-EVRAG requires the least time, increasing from approximately 16 ms to 29.5 ms. The proposed CLR-EVRAG performs better than the other two methods at every image size. This result shows that it can process fundus images with less computational time. Even when the number of images reaches 10,000, the proposed method maintains a lower processing time than CNN-RAG and Retina-RAG. Therefore, CLR-EVRAG provides better computational efficiency and is suitable for handling large numbers of fundus images.

4.4 Test Case 4: False Positive Rate

The false positive rate measures the proportion of actual negative cases that are incorrectly classified as positive by a model. In the proposed CLR-EVRAG framework, it indicates how often retinal images without a particular diabetic retinopathy (DR) condition or lesion are incorrectly identified as positive. A lower false positive rate indicates that the model makes fewer incorrect positive predictions.

Table 4 Simulation Performance of False Positive Rate

Number Of Fundus Images	False Positive Rate (%)		
	CNN-RAG	Retina-RAG	Proposed CLR-EVRAG
2000	6.95	4.55	2.94
4000	6.58	4.26	2.52
6000	6.26	3.92	2.07
8000	6.09	3.38	1.88
10000	5.77	2.73	1.36

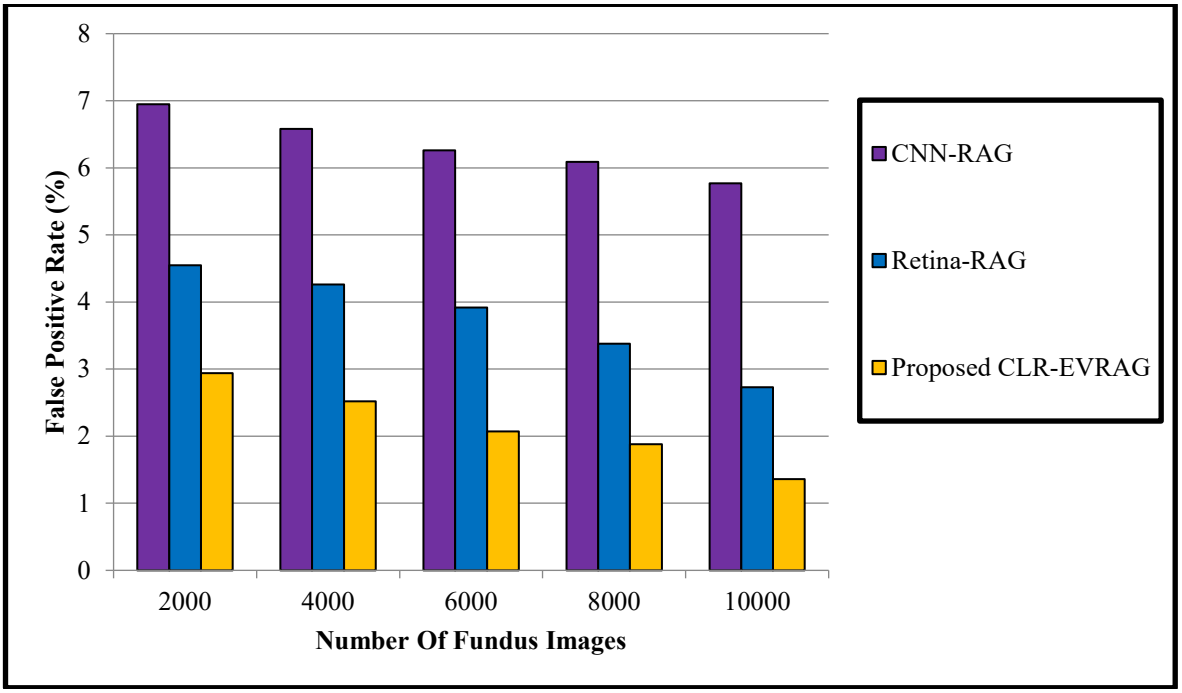


Figure 5 Graphical Representation of False Positive Rate

Table 4 and Figure 5 shows the false positive rate (FPR) of three methods: CNN-RAG, Retina-RAG, and the proposed CLR-EVRAG method using different numbers of fundus images. The false positive rate is measured as a percentage. A lower value indicates that the method produces fewer false positive results. As shown in the graph, the false positive rate decreases as the number of fundus images increases for all three methods. CNN-RAG [1] shows the highest rate, decreasing from approximately 6.9% at 2,000 images to 5.8% at 10,000 images. Retina-RAG [2] performs better than CNN-RAG, with its rate falling from about 4.5% to 2.7%. In comparison, the proposed CLR-EVRAG achieves the lowest false positive rate, reducing it from approximately 2.9% to 1.4%. The proposed CLR-EVRAG performs better than the other two methods across all tested image sizes. Its lower false positive rate indicates that it can reduce incorrect positive predictions during fundus image analysis. This improvement can help provide more reliable classification results. Overall, the graph shows that the proposed CLR-EVRAG offers better performance in controlling false positive results than CNN-RAG [1] and Retina-RAG [2].

5. CONCLUSION

This paper presented CLR-EVRAG with aiming at enhancing the explainability and reliability of DR severity classification and CRG. The CLR-EVRAG integrates retinal lesion detection, severity classification, counterfactual reasoning, clinical evidence retrieval, and report verification within a unified process. Lesion detection model in CLR-EVRAG predicts key retinal abnormalities, while CLCON based severity classification module predicts the appropriate DR grade. Counterfactual reasoning in CLR-EVRAG examines how changes in selected lesion-related information influence the model's prediction. The RAG module in CLR-EVRAG mines relevant clinical information from trusted sources to support the explanation of the predicted results. Finally, the evidence verification step checks whether the statements in the generated report agree with the detected lesions, classification results, and retrieved clinical references. Through verifying for unsupported statements, conflicting findings, and uncertain predictions, the CLR-EVRAG aims to decrease misleading explanations and enhances report consistency. Poor-quality images and insufficient evidence can be flagged for further review rather than being used to produce overly confident conclusions. Overall, CLR-EVRAG provides a structured approach for designing more transparent, evidence-supported, and clinically interpretable DR assessment systems. Thus, CLR-EVRAG assist healthcare professionals in understanding model predictions rather than replace ophthalmologists or independent clinical diagnosis.

REFERENCES

[1]Alshehri, Afrah, and Arwa Basbrain. "Automated Diabetic Retinopathy Severity Assessment and Narrative Reporting via CNN Classification and Retrieval-Augmented Generation (RAG)." In 2026 Asia Conference on Computer Vision, Image Processing and Pattern Recognition (CVIPPR), pp. 1-10. IEEE, 2026.

[2] Zaian, Abdelrahman, Sheethal Bhat, Mohamed Abdalkader, and Andreas Maier. "Retina-RAG: Retrieval-Augmented Vision-Language Modeling for Joint Retinal Diagnosis and Clinical Report Generation." arXiv preprint arXiv:2605.06173 (2026).

[3] Ito, Masato, Kaito Tanaka, Keisuke Matsuda, and Aya Nakayama. "XDR-LVLM: An explainable vision-language large model for diabetic retinopathy diagnosis." arXiv preprint arXiv:2508.15168 (2025).

[4] Agarwal, Ambuj Kumar, Abu Bakar Bin Abdul Hamid, Danish Ather, Raj Gaurang Tiwari, Indrajit De, and Kunchanapaalli Rama Krishna. "Lesion-aware ordinal transformer for diabetic retinopathy classification from fundus images." Biomedical and Pharmacology Journal 19, no. 1 (2026): 233-245.

[5] Akhtar, Samia, Shabib Aftab, Oualid Ali, Munir Ahmad, Muhammad Adnan Khan, Sagheer Abbas, and Taher M. Ghazal. "A deep learning based model for diabetic retinopathy grading." Scientific reports 15, no. 1 (2025): 3763.
[6] Ikram, Amna, and Azhar Imran. "ResViT FusionNet Model: An explainable AI-driven approach for automated grading of diabetic retinopathy in retinal images." Computers in biology and medicine 186 (2025): 109656.
[7] Alamri, Faten S., Tariq Sadad, Kolawole John Adebayo, Abeer Rashad Mirdad, and Amjad Rehman. "Ensemble of graph neural networks with Transformer fusion for diabetic retinopathy severity classification." Egyptian Informatics Journal 34 (2026): 100946.
[8] Aftab, Shabib, and Samia Akhtar. "Diabetic retinopathy severity classification using data fusion and ensemble transfer learning." Journal of Software Engineering and Applications 18, no. 1 (2025): 1-23.
[9] Saleem, Sehrish, Ramzan Talib, Muhammad Kashif Hanif, and Muhammad Umer Sarwar. "Clinical-guided deep learning framework for diabetic retinopathy: integrating lesion-aware attention, adversarial augmentation, and uncertainty quantification." Scientific Reports (2026).
[10] Rezaee, Khosro, and Fateme Farnami. "Innovative approach for diabetic retinopathy severity classification: An ai-powered tool using cnn-transformer fusion." Journal of Biomedical Physics & Engineering 15, no. 2 (2025): 137.
[11] Yan, Xin, Shiqi Lei, Lifen Hu, Mu Qin, and Na Wu. "Diagnostic accuracy and clinical performance of deep learning models for grading diabetic retinopathy: a systematic review and meta-analysis." Frontiers in Endocrinology 17 (2026): 1853785.
[12] Swapna, M., Tanishqa Ravirala, and Nikhitha Reddy. "Diabetic retinopathy detection using federated learning and vision transformers." International Journal of Interpreting Enigma Engineers (IJIEE) 2, no. 1 (2025): 10-21.
[13] Yadav, Uma, Palavi Meshram, Shweta Bondre, Vipin Bondre, and Priya Dasarwar. "Deep Learning-Based Classification of Diabetic Retinopathy Using Fundus Images and Transfer Learning." In 2025 Second International Conference on Pioneering Developments in Computer Science & Digital Technologies (IC2SDT), pp. 220-225. IEEE, 2025.
[14] Vohra, Harpreet, Mohammad Kamrul Hasan, Siti Norul Huda Sheikh Abdullah, Shayla Islam, Abdul Hadi Abd Rahman, Shreya N. Bhojake, Dhruv Bobal, Awal Kaur Gandhi, and Chandramohan Dhasarathan. "A low-cost ai-powered system for early detection of diabetic retinopathy and ocular diseases in resource limited settings." IEEE Access 13 (2025): 97322-97336.
[15] Zhang, Xiaoyun, Xianjun Li, Xinpan Yuan, Pinjie He, Mingming Wang, and Jue Wang. "IRG-DR: Interactive Report Generation Framework with Evidence Guidance for Diabetic Retinopathy." In 2026 29th International Conference on Computer Supported Cooperative Work in Design (CSCWD), pp. 2791-2796. IEEE, 2026.
[16] Huang, Jia-Hong. "DeepEyeNet: Generating medical report for retinal images." In Proceedings of the 34th ACM International Conference on Information and Knowledge Management, pp. 6789-6792. 2025.
[17] Palaniappan, Damodharan, Tan Kuan Tak, K. Vijayan, Balajee Maram, Pravin R. Kshirsagar, and Naim Ahmad. "Enhancement of medical imaging technique for diabetic retinopathy: Realistic synthetic image generation using GenAI." Computer Modeling in Engineering & Sciences 145, no. 3 (2025): 4107-4127.
[18] Gundu, Aakanksha, Nihar Vaidya, Peter Jones, and Ann Ostrovsky. "RAG to Riches: The Effectiveness of Retrieval-Augmented Generation for Diabetic Retinopathy Classification." Investigative Ophthalmology & Visual Science 67, no. 7 (2026): 3181-3181.
[19] Zhou, Jiaying, Shuyi Zhang, Xiaoyan Hu, Bowen Xu, Miles Wong, Simon Szeto, and Carol Y. Cheung. "A Comparative Evaluation of Multi-Agent Systems With and Without Specialized Image-based Tools for Diabetic Retinal Disease Detection and Clinical Report Generation." Investigative Ophthalmology & Visual Science 67, no. 7 (2026): 3371-3371.
[20] Han, Xiaoyan, Jin Ma, Jinghan Zhang, Kunpeng Liu, and Feng Luo. "Understanding the Constraints of RAG-Based Medical LVLMs-A Case Study in Ophthalmic Report Generation." In 2025 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 96-102. IEEE, 2025.