Research Square

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

DRNet++: A Lesion-Aware Hybrid ConvNeXtV2--Vision Transformer Framework for Diabetic Retinopathy Grading and Weakly Supervised Lesion Localization

Shikha Singh
Bennett University
Nitin Arvind Shelke
nitin.shelke@bennett.edu.in
Bennett University
Manoj K. Singh
Bennett University

Research Article

Keywords: Clinical Decision Support, ConvNeXtV2, Diabetic Retinopathy, Explainable Artificial Intelligence, Vision Transformer, Weakly Supervised Lesion Localization.

Posted Date: November 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10361047/v1

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

Additional Declarations: No competing interests reported.

DRNet++: A Lesion-Aware Hybrid ConvNeXtV2–Vision Transformer Framework for Diabetic Retinopathy Grading and Weakly Supervised Lesion Localization

Shikha Singh[1*], Nitin Arvind Shelke[1*†] and Manoj K. Singh[1*†]

[1*]School of computer science, engineering, and technology, Bennett University, Greater Noida, Uttar Pradesh, India.

*Corresponding author(s). E-mail(s): e22soep0052@bennett.edu.in; nitin.shelke@bennett.edu.in; manoj.kumar4@bennett.edu.in;
†These authors contributed equally to this work.

Abstract

Diabetic Retinopathy (DR) is one of the leading causes of preventable vision loss worldwide, necessitating accurate and interpretable automated screening systems for timely diagnosis. Although recent deep learning methods have achieved impressive classification performance, most existing approaches either emphasize local lesion representation while overlooking global contextual relationships or employ transformer-based architectures that inadequately preserve subtle retinal lesion characteristics. Moreover, limited attention has been devoted to integrating lesion localization with disease classification under weakly supervised settings. To overcome these challenges, we propose **DRNet++**, a hybrid CNN–Transformer framework for diabetic retinopathy grading and weakly supervised lesion localization. The proposed framework integrates four key components, namely the Lesion Evidence Enhancement Module (LEEM), Evidence-Guided Token Generation (EGTG), Clinical Context Transformer (CCT), and Dynamic Lesion Evidence Fusion (DLEF), to jointly enhance lesion-aware feature extraction, global contextual modeling, and adaptive feature fusion. Extensive experiments conducted on the publicly available DDR dataset demonstrate that DRNet++ achieves an accuracy of **95.93%**, precision of **96.07%**, recall of **95.70%**, F1-score of **95.88%**, AUC of **97.52%**, and Cohen's $\boldsymbol{\kappa}$ of **0.913**, consistently outperforming representative CNN-, transformer-, and hybrid-based methods. Furthermore, evaluation on the IDRiD dataset using Grad-CAM++ demonstrates effective weakly supervised lesion localization, achieving a Dice coefficient

of **0.69** and overall IoU of **0.61** without requiring pixel-level lesion annotations during training. These results demonstrate that DRNet++ provides an interpretable, accurate and clinically applicable framework for automated diabetic retinopathy screening.

Keywords: Clinical Decision Support, ConvNeXtV2, Diabetic Retinopathy, Explainable Artificial Intelligence, Vision Transformer, Weakly Supervised Lesion Localization.

1 Introduction

Diabetic Retinopathy (DR) is the most common microvascular complication of diabetes mellitus and a leading cause of irreversible visual loss and blindness in working-age people throughout the world [1, 2]. Persistent hyperglycemia leads to insidious microvascular injury in the retinal microvasculature which eventually translates to cotton wool spots,hard exudates, microaneurysms, hemorrhages and neovascularization. Delay in early detection and management leads to irreversible blindness due to visually disabling diabetic retinopathy. Recent epidemiological studies continue to report a growing prevalence of diabetic retinopathy, emphasizing the urgent need for scalable, accurate, and cost-effective screening systems capable of supporting early diagnosis and timely clinical intervention [3].

Retinal fundus imaging is widely employed for diabetic retinopathy screening because it is non-invasive and provides clinically valuable information about retinal abnormalities. Nevertheless, manual interpretation of color fundus photographs by ophthalmologists is labor-intensive, time-consuming, and prone to inter-observer variability, making large-scale screening particularly challenging in resource-limited settings. To overcome this challenge, AI, particularly deep learning algorithms, presents an efficient alternative and shows a high potential to reach the accuracy comparable to human experts for the screening and detection of diabetic retinopathy [4, 5]. The studies by Ting *et al.* [6] and Gulshan *et al.* [7] validated that deep convolutional neural networks (DCNN) can achieve high sensitivity and specificity in large-scale diabetic retinopathy screening, paving the way for AI-assisted ophthalmic diagnosis.

With the popularity of CNNs, there has been a surge in research into the automated detection, grading, lesion detection and lesion segmentation of diabetic retinopathy (DR). Numerous CNN-based architectures have been proposed by incorporating transfer learning, attention mechanisms, multi-scale feature extraction, and class imbalance handling to improve diagnostic performance [8–13]. Similarly, lesion-aware segmentation methods have focused on accurately localizing pathological regions, including microaneurysms, hemorrhages, optic disc, and diabetic macular edema, thereby providing valuable structural information for disease grading [14–20]. Comprehensive surveys further indicate that CNN-based models have substantially advanced diabetic retinopathy analysis while highlighting persistent challenges related to lesion diversity, class imbalance, image quality variations, and limited model interpretability [21–23]. Despite these advances,

Although CNNs effectively capture local spatial patterns through convolutional operations, their accessible fields often limit the modeling of long-range contextual dependencies that are essential for distinguishing subtle retinal abnormalities distributed across different anatomical regions. To address this limitation, Vision Transformers (ViTs) have recently attracted significant attention by using self-attention mechanisms to model global feature interactions within retinal images. Inspired by the original Vision Transformer architecture [24], several transformer-based diabetic retinopathy frameworks have demonstrated superior representation learning and improved grading accuracy by exploiting hierarchical attention, multi-scale feature aggregation, and lesion-aware feature encoding [25–34]. Hybrid CNN-Transformer architectures further combine the strong local feature extraction capability of CNNs with the global contextual reasoning of transformers, resulting in improved robustness and grading performance [35–39].

These efforts make considerable progress, yet many fundamental research issues have yet to be solved. For example, traditional CNN-based models generally cannot sufficiently consider the global contextual relationships between several lesions at the fine-grained level. Transformer-based models generally require large-scale pretraining or sufficient annotated data to achieve optimal performance. Moreover, diabetic retinopathy datasets typically exhibit severe class imbalance, heterogeneous lesion distributions, varying illumination conditions, and substantial inter-patient variability, which often reduce model generalization across diverse clinical settings [32, 40, 41]. Another important limitation of existing models is basically the lack of explainability in many deep learning frameworks, making clinical validation and physician trust difficult despite high predictive accuracy. Although methods such as Grad-CAM and concept-aware attention have improved model interpretability, generating clinically meaningful visual explanations while maintaining high classification performance remains an active research problem [37, 42, 43]. Recent studies have further explored self-supervised learning, graph neural networks, adaptive feature fusion, and explainable transformer architectures to improve robustness and interpretability in diabetic retinopathy diagnosis [44–50]. Nevertheless, achieving an optimal balance among classification accuracy, computational efficiency, robustness, lesion awareness, and model interpretability continues to be a challenging research objective, particularly for deployment in real-world screening environments.

Unlike existing CNN, transformer, and hybrid approaches, DRNet++ jointly integrates lesion evidence enhancement, evidence-guided token generation, contextual transformer reasoning, and adaptive feature fusion within a unified end-to-end framework while simultaneously providing weakly supervised lesion localization. To address these limitations, this paper proposes **DRNet++**, a hybrid CNN–Transformer framework for diabetic retinopathy severity classification and weakly supervised lesion localization. The framework combines convolutional feature extraction with transformer-based contextual modeling to effectively learn both local lesion characteristics and global retinal dependencies. Furthermore, the framework incorporates lesion-aware representation learning and adaptive feature fusion to improve disease grading while generating clinically interpretable lesion localization maps

through Grad-CAM++ without requiring pixel-level lesion annotations. The main contributions of this work are given below:

- A novel hybrid CNN–Transformer framework, namely **DRNet++**, is proposed for diabetic retinopathy severity classification by effectively integrating local lesion representations with global contextual information.
- A *Lesion Evidence Enhancement Module (LEEM)* is proposed to enhance lesion-sensitive feature representations by jointly exploiting semantic and spatial lesion evidence.
- An *Evidence-Guided Token Generation (EGTG)* module, together with a *Clinical Context Transformer (CCT)* and a *Dynamic Lesion Evidence Fusion (DLEF)* module, is introduced to capture complementary local and global contextual information for robust disease grading.
- A weakly supervised lesion localization strategy based on Grad-CAM++ is employed to provide clinically interpretable visual explanations without requiring pixel-level lesion annotations during training.
- Extensive experiments on the DDR and IDRiD benchmark datasets demonstrate superior classification performance, improved lesion localization capability, and strong cross-dataset generalization compared with representative CNN-based, transformer-based, and recent state-of-the-art methods.

This study is organized as follows: Section 2 summarizes the relevant literature. Section 3 introduces the proposed DRNet++ framework. Section 4 discusses the experimental results, and Section 5 provides the conclusion and future direction.

2 Related Work

Due to its potential for widespread screening and early disease detection, automated diabetic retinopathy (DR) grading has emerged as one of the most actively researched uses of deep learning in ophthalmic image analysis. Deep learning's effectiveness has reduced reliance on manually created picture attributes while greatly increasing the accuracy of eye disease identification. Several comprehensive surveys have summarized recent advances in DR diagnosis and highlighted the growing importance of robust, interpretable, and computationally efficient artificial intelligence systems for clinical deployment [21–23]. However, developing models with simultaneous excellent disease grading performance, adequate lesion modeling, accurate global contextual reasoning and clinically meaningful interpretability is still an open research problem. The early deep learning-based methods for diabetic retinopathy severity grading mostly adopt the CNN method because of their superb ability of hierarchical learning of image representation directly from the retinal fundus images. CNN-based systems can achieve ophthalmologist-level performance for automated DR screening that is demonstrated in Large-scale clinical studies [6, 7, 51]. Subsequent studies further improved classification accuracy by employing transfer learning, optimized CNN architectures, attention mechanisms, and feature enhancement strategies [10–13, 41, 52, 53]. CNN-based approaches effectively capture local retinal abnormalities such as hemorrhages, microaneurysms, and exudates through hierarchical convolutional feature extraction.

However, their localized receptive fields limit their ability to model long-range spatial connection among lesions dispersed over different retinal regions. Furthermore, repeated pooling operations may suppress subtle pathological structures that are essential for accurate early-stage diabetic retinopathy grading.

As transformer-based architectures may simulate long-range relationships utilizing self-attention methods, they have garnered a lot of attention recently as a way to get around these restrictions. By capturing global contextual relationships throughout the entire fundus image, Vision Transformer (ViT) and hierarchical transformer variations have shown a promising outcome for retinal image analysis [25–27, 29–31, 33, 54]. More recent transformer-based frameworks further incorporate lesion-aware attention and hierarchical feature aggregation to improve discrimination among different diabetic retinopathy severity levels [34, 37, 38]. Although these approaches achieve improved contextual modeling, pure transformer architectures generally require large-scale annotated datasets, involve higher computational complexity, and may overlook fine-grained lesion characteristics because of their patch-based representation strategy.

Recent research has increasingly concentrated on hybrid CNN–Transformer architectures that jointly take use of local texture representation and global contextual reasoning, acknowledging the complimentary benefits of CNNs and transformers. Representative studies combine convolutional feature extraction with transformer-based attention modules, multi-scale feature fusion, or cross-attention mechanisms to improve diabetic retinopathy grading performance [19, 28, 35, 45]. These hybrid frameworks consistently outperform conventional CNN-only and transformer-only models by integrating local lesion information with global semantic representations. Nevertheless, existing approaches mainly emphasize feature fusion after backbone feature extraction and seldom explicitly enhance lesion evidence before token generation or adaptively integrate clinically relevant lesion representations throughout the learning process. Consequently, subtle retinal abnormalities may still receive insufficient attention during feature learning.

Beyond improving classification accuracy, explainability has become an essential requirement for trustworthy clinical decision support systems. To improve model transparency and clinician confidence, gradient-based visualisation techniques like Grad-CAM and Grad-CAM++ have been widely used to highlight image regions that contribute most significantly to model predictions [42]. More recently, several studies have explored lesion-aware learning and weakly supervised localization to reduce dependence on expensive pixel-level annotations while improving model interpretability [8, 17, 19, 20, 55, 56]. Although these approaches successfully identify clinically relevant retinal regions, lesion localization is typically performed as a post-hoc explanation and is rarely incorporated into the feature representation learning process. As a result, the interaction between lesion localization and disease classification remains insufficiently explored in existing diabetic retinopathy grading frameworks. For clarity, a summary of representative diabetic retinopathy classification methods is presented in Table 1. As evident from the comparison, existing studies have primarily focused on improving classification accuracy or interpretability independently, whereas the unified integration of lesion enhancement, contextual feature learning, adaptive feature fusion, and weakly supervised lesion localization remains largely unexplored.

Table 1: Summary of representative methods for diabetic retinopathy analysis.

Reference	Method / Model	Dataset	Explainability	Limitation
Gulshan *et al.* [7]	Deep CNN	EyePACS	No	Focuses only on disease classification without lesion localization or model interpretability.
Ting *et al.* [6]	CNN-based screening system	Multi-ethnic retinal dataset	No	Requires large-scale training data and provides limited explanation of predictions.
Lin and Wu [10]	Modified ResNet-50	DDR	No	Limited capability to capture long-range contextual dependencies.
Gu *et al.* [26]	Vision Transformer with Residual Attention	DDR	Partial	Global attention may overlook subtle local retinal lesions.
Wang *et al.* [45]	Graph Convolution Network with Multi-scale Fusion	DDR	No	Computationally complex and lacks lesion localization capability.
Zhang *et al.* [19]	Multi-Lesion Segmentation Guided Attention Network	DDR	Yes	Relies on lesion-guided supervision, increasing annotation requirements.
Bala *et al.* [28]	CNN–Transformer Network (CTNet)	APTOS 2019	No	Does not explicitly enhance lesion evidence before feature fusion.
Sadeghzadeh *et al.* [35]	Hybrid CNN–Transformer	APTOS 2019	No	Limited adaptive fusion between local and global representations.
Wen *et al.* [37]	Concept-based Lesion-aware Transformer	Private retinal dataset	Yes	Emphasizes explainability but does not perform unified DR grading and localization.
Zhang *et al.* [56]	CNN, ViT and Hybrid Comparative Framework	APTOS, EyePACS	Yes	Comparative study without proposing a unified diagnostic framework.

Overall, the existing literature demonstrates substantial progress in automated diabetic retinopathy grading using CNN-based, transformer-based, and hybrid deep learning architectures. Nevertheless, existing methods rarely integrate lesion enhancement, lesion-aware token generation, clinical context modeling, adaptive feature fusion, and weakly supervised lesion localization within a unified framework. Motivated by these limitations, the proposed DRNet++ framework addresses these challenges through an end-to-end hybrid ConvNeXtV2–Vision Transformer architecture that jointly improves classification performance and model interpretability.

3 Proposed Methodology

This section presents the proposed DRNet++ framework for automated DR grading and weakly supervised lesion localization. First, the problem is formally defined, followed by the overall network architecture and the proposed modules. Finally, the optimization objective and training strategy are described.

3.1 Problem Formulation

Let the training dataset be denoted as

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}, \tag{1}$$

where $\mathbf{x}_i \in \mathbb{R}^{H\times W\times 3}$ represents the i^{th} retinal fundus image, $y_i \in \{0, 1, \ldots, K-1\}$ denotes the corresponding diabetic retinopathy (DR) severity label, N is the total

number of training images, and K represents the number of DR severity categories. Since only image-level diagnostic labels are available during training, pixel-level lesion annotations are not utilized for network optimization. Consequently, the proposed approach is formulated as a weakly supervised learning problem.

Given an input retinal image $\mathbf{x}$, the objective is to learn a nonlinear mapping

$$f_\theta : \mathbf{x} \rightarrow (\hat{y}, \mathbf{E}), \tag{2}$$

where θ denotes the trainable model parameters, $\hat{y}$ is the predicted DR severity class, and $\mathbf{E}$ represents the learned lesion evidence representation. Unlike conventional image classification models that directly optimize class predictions, the proposed approach aims to learn discriminative lesion-aware representations capable of preserving subtle pathological structures while simultaneously capturing their global contextual relationships.

Based on the learned lesion evidence representation, a lesion localization function is defined as

$$\mathbf{L} = g(\mathbf{E}, \hat{y}), \tag{3}$$

where $\mathbf{L} \in \mathbb{R}^{H \times W}$ denotes the generated lesion localization map and $g(\cdot)$ represents the localization function. Since lesion annotations are unavailable during training, the localization process is learned implicitly from image-level supervision and therefore follows the weakly supervised learning paradigm.

The overall optimization objective is formulated as

$$\theta^* = \arg\min_\theta \mathcal{L}_{\text{total}}, \tag{4}$$

where $\mathcal{L}_{\text{total}}$ denotes the overall learning objective used to optimize proposed framework. The detailed formulation of the optimization strategy and its individual loss components are presented in Section 3.9. The mathematical symbols and notations employed in the subsequent methodological descriptions are summarized in Table 2.

3.2 Overview of the Proposed Framework

The proposed framework is designed to perform diabetic retinopathy (DR) severity classification and weakly supervised lesion localization simultaneously within a unified end-to-end learning framework. Unlike conventional DR classification methods that directly learn image-level representations, the proposed approach explicitly learns lesion evidence representations that preserve subtle pathological characteristics while capturing global contextual dependencies among retinal abnormalities. This enables the framework to produce accurate disease predictions together with clinically interpretable lesion localization.

The suggested framework, as shown in Fig. 1, consists of the six sequential stages: (i) retinal fundus image pre-processing, (ii) lesion-aware hierarchical feature extraction, (iii) lesion evidence enhancement, (iv) evidence-guided contextual representation learning, (v) DLEF, and (vi) weakly supervised lesion localization. Initially, the input retinal image undergoes a lightweight pre-processing step to reduce appearance variations while preserving diagnostically relevant retinal structures. Hierarchical feature

Table 2: Summary of notations used in the proposed framework.

Notation	Description
$\mathcal{D}$	Training dataset
N	Number of training samples
$\mathbf{x}$	Input retinal fundus image
y	Ground-truth DR severity label
$\hat{y}$	Predicted DR severity label
K	Number of DR severity classes
$\mathbf{F}$	Hierarchical feature representation
$\mathbf{F}^e$	Lesion-enhanced feature representation
$\mathbf{T}^e$	Lesion-aware token sequence
$\mathbf{Z}$	Contextual token representation
$\mathbf{E}$	Unified lesion evidence representation
$\mathbf{L}$	Lesion localization map
$\mathbf{M}_e$	Lesion evidence map generated by LEEM
$\mathcal{L}_{total}$	Overall optimization loss

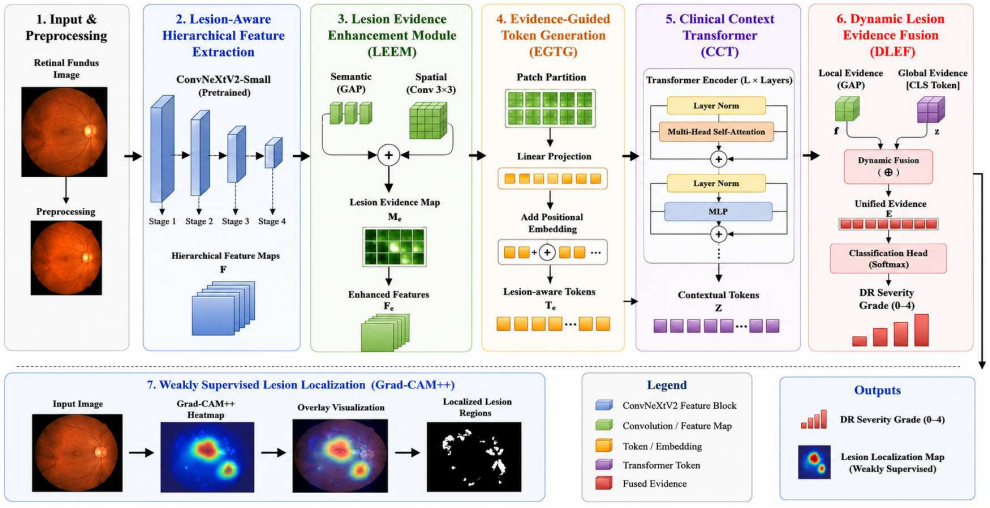


Fig. 1: Proposed DRNet++ framework

representations are subsequently extracted to capture both fine-grained lesion characteristics and high-level semantic information. The extracted features are further refined through a lesion evidence enhancement module to reduce irrelevant retinal responses and highlight lesion-sensitive regions.

The refined lesion representations are transformed into evidence-guided contextual embeddings that model long-range relationships among spatially distributed retinal lesions. The complementary local and global representations are then adaptively fused to construct a unified lesion evidence representation for DR severity prediction.

Finally, the learned lesion evidence is utilized to generate class-discriminative lesion localization maps without requiring pixel-level lesion annotations during training, thereby preserving the weakly supervised learning paradigm. The detailed architecture and mathematical formulation of each component are presented in the following subsections.

3.3 Lesion-Aware Hierarchical Feature Extraction

Accurate diabetic retinopathy (DR) severity assessment requires the identification of retinal lesions exhibiting significant variations in shape, size, texture, and spatial distribution. Early-stage pathological signs, such as microaneurysms, often occupy only a few pixels, whereas advanced lesions, including hemorrhages and hard exudates, may extend over relatively larger retinal regions. Consequently, an effective feature extraction network should preserve subtle lesion characteristics while simultaneously learning high-level semantic representations capable of distinguishing different DR severity stages.

Given a pre-processed retinal fundus image $\mathbf{x}$, the proposed architecture first extracts hierarchical feature representations using a multi-stage convolutional backbone. Unlike conventional convolutional networks that primarily emphasize high-level semantic information, hierarchical feature learning progressively captures both low-level anatomical structures and lesion-specific characteristics across multiple spatial resolutions. This assists the network to retain fine-grained pathological details while increasing receptive field for robust disease representation.

The hierarchical feature extraction process is formulated as

$$\mathbf{F} = F_{\mathrm{H}}(\mathbf{x}), \tag{5}$$

where $F_{\mathrm{H}}(\cdot)$ denotes the hierarchical feature extraction function and $\mathbf{F} \in \mathbb{R}^{H_f \times W_f \times C}$ represents the resulting feature maps, with H_f, W_f, and C denoting the spatial dimensions and channel depth, respectively.

In this work, the hierarchical feature extractor is instantiated using the ConvNeXtV2-Small architecture owing to its strong capability for learning multi-scale feature representations, large effective receptive fields, and computational efficiency. The extracted feature maps jointly encode retinal anatomical structures together with lesion-sensitive responses, thereby providing a comprehensive representation for subsequent lesion evidence learning.

Although the hierarchical representations contain abundant diagnostic information, they may also include activations corresponding to normal retinal tissues and background structures that contribute little to disease discrimination. Therefore, the extracted feature maps are further refined by the proposed LEEM, which selectively emphasizes clinically relevant lesion regions while suppressing redundant feature responses.

3.4 Lesion Evidence Enhancement Module

Although the hierarchical feature extractor learns rich multi-scale retinal representations, the extracted feature maps often contain redundant responses originating

from normal anatomical structures, such as optic disc boundaries, retinal vessels, and homogeneous background regions. These responses may dominate the learned feature distribution and obscure subtle pathological patterns, particularly during the early stages of diabetic retinopathy (DR), where lesion regions occupy only a small fraction of retinal image. Consequently, directly utilizing the extracted feature maps may reduce the discriminative capability of the subsequent contextual representation learning module.

To address this challenge, the proposed framework introduces a *LEEM* that explicitly enhances lesion-sensitive representations before global contextual modeling. Unlike conventional attention mechanisms that independently estimate channel-wise and spatial attention, LEEM jointly learns semantic and spatial lesion evidence to construct a unified lesion evidence map. This design enables the network to emphasize diagnostically relevant pathological regions while preserving the underlying anatomical information through residual feature enhancement.

Let the hierarchical feature representation extracted from the backbone network be denoted as

$$\mathbf{F} \in \mathbb{R}^{H_f \times W_f \times C}, \tag{6}$$

where H_f, W_f, and C denote the spatial dimensions and the number of feature channels, respectively.

First, global semantic evidence is extracted by aggregating channel-wise contextual information using Global Average Pooling (GAP), followed by a 1×1 convolutional projection,

$$\mathbf{S} = \mathrm{Conv}_{1\times1}\left(\mathrm{GAP}(\mathbf{F})\right), \tag{7}$$

where $\mathbf{S}$ represents the semantic lesion evidence describing the relative importance of discriminative retinal features.

In parallel, local spatial evidence is obtained by applying a 3×3 convolution over the feature maps,

$$\mathbf{P} = \mathrm{Conv}_{3\times3}(\mathbf{F}), \tag{8}$$

where $\mathbf{P}$ captures the spatial distribution of lesion-sensitive regions while preserving local structural continuity.

The semantic and spatial evidence representations are subsequently integrated to estimate the unified lesion evidence map,

$$\mathbf{M}_e = \sigma\left(\mathbf{S} + \mathbf{P}\right), \tag{9}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function that normalizes evidence scores into the interval $[0, 1]$.

Finally, the estimated lesion evidence map is incorporated into the original feature representation using a residual enhancement strategy,

$$\mathbf{F}^e = \mathbf{F} + \mathbf{M}_e \odot \mathbf{F}, \tag{10}$$

where $\odot$ denotes element-wise multiplication and $\mathbf{F}^e$ represents the final lesion-enhanced feature representation.

By jointly exploiting semantic and spatial lesion evidence, LEEM effectively suppresses redundant retinal background responses while preserving subtle pathological structures, including microaneurysms, hemorrhages, and exudates. The enhanced feature representation provides a robust and discriminative feature space for subsequent global contextual representation learning, thereby facilitating more reliable DR severity classification under weakly supervised learning.

3.5 Evidence-Guided Token Generation

Although the LEEM effectively emphasizes diagnostically relevant retinal regions, the resulting feature representation remains in a two-dimensional spatial form. However, transformer-based architectures operate on sequential token representations and therefore require the feature maps to be converted into a sequence of informative embeddings. Conventional Vision Transformers perform uniform patch tokenization, assigning equal importance to all image patches regardless of their clinical relevance. Such a strategy may introduce redundant background information and reduce the contribution of subtle lesion regions during contextual representation learning.

To overcome this limitation, the proposed architecture introduces an $EGTG$ module that transforms the lesion-enhanced feature representation into a sequence of lesion-aware tokens. Since the input feature maps have already been refined by LEEM, each generated token inherently preserves discriminative pathological evidence while suppressing irrelevant retinal structures.

Let the lesion-enhanced feature representation be denoted by

$$\mathbf{F}^e \in \mathbb{R}^{H_f \times W_f \times C}, \tag{11}$$

where H_f, W_f, and C represent the spatial dimensions and channel depth, respectively. The feature maps are first partitioned into a set of non-overlapping feature patches, each of which is projected into a fixed-dimensional embedding space using a linear projection function,

$$\mathbf{T} = P(\mathbf{F}^e), \tag{12}$$

where $P(\cdot)$ denotes the patch embedding operation and

$$\mathbf{T} = \left\{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_{N_p}\right\} \in \mathbb{R}^{N_p \times d},$$

with N_p representing the number of generated tokens and d denoting the embedding dimension.

To preserve the spatial correspondence among retinal regions, learnable positional embeddings are incorporated into the generated token sequence,

$$\mathbf{T}^e = \mathbf{T} + \mathbf{E}_{pos}, \tag{13}$$

where $\mathbf{E}_{pos}$ denotes the learnable positional embedding matrix.

Unlike conventional patch tokenization performed directly on the input image, the proposed EGTG module generates tokens from lesion-enhanced feature maps. Consequently, each token encodes both local pathological characteristics and lesion-sensitive semantic information learned during the preceding feature enhancement stage. This

lesion-aware token representation provides a more informative input for transformer-based contextual modeling while substantially reducing the influence of redundant retinal background information.

The generated token sequence $\mathbf{T}^e$ is subsequently forwarded to the Clinical Context Transformer to model long-range dependencies among spatially distributed retinal lesions.

3.6 Clinical Context Transformer

Diabetic retinopathy severity is determined not only by the presence of individual retinal lesions but also by their spatial distribution and contextual relationships across the retina. Although convolutional feature extraction effectively captures local pathological characteristics, its inherently localized receptive field limits the modeling of long-range dependencies among anatomically distant lesion regions. Consequently, relying solely on convolutional representations may lead to incomplete characterization of disease severity, particularly in advanced stages where multiple lesion types coexist. To overcome this problem, the proposed approach includes a *EGTG* that performs global contextual reasoning over the lesion-aware tokens generated by the EGTG module. By allowing each lesion-aware token to interact with every other token, the transformer learns complementary contextual relationships that cannot be effectively captured through convolution alone.

Given the lesion-aware token sequence

$$\mathbf{T}^e = \left\{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_{N_p}\right\}, \tag{14}$$

the transformer encoder learns contextual representations according to

$$\mathbf{Z} = T_{\text{enc}}(\mathbf{T}^e), \tag{15}$$

where $T_{\text{enc}}(\cdot)$ denotes the transformer encoder and

$$\mathbf{Z} = \left\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{N_p}\right\} \in \mathbb{R}^{N_p \times d}$$

represents the contextualized lesion token representations.

Within the transformer encoder, the self-attention mechanism enables each token to selectively aggregate information from all other lesion-aware tokens, thereby capturing global interactions among spatially separated retinal abnormalities. Consequently, pathological patterns distributed across different retinal regions can jointly contribute to the learned feature representation, facilitating a more comprehensive understanding of disease severity.

Unlike conventional Vision Transformers that operate directly on image patches, the proposed CCT receives lesion-aware embeddings generated by the EGTG module. Therefore, the transformer focuses predominantly on clinically relevant retinal information rather than uniformly processing both lesion and background regions. This lesion-centric contextual reasoning improves the discriminative capability of the learned representations while preserving computational efficiency.

The contextual lesion representations generated by the CCT are subsequently integrated with the local lesion evidence through the proposed DLEF module for final diabetic retinopathy severity prediction.

3.7 Dynamic Lesion Evidence Fusion

The hierarchical convolutional features and transformer-based contextual representations provide complementary descriptions of retinal pathology. While the lesion-enhanced convolutional features preserve fine-grained local characteristics of retinal abnormalities, the Clinical Context Transformer captures long-range contextual dependencies among spatially distributed lesions. Simply concatenating these heterogeneous representations may not fully exploit their complementary strengths, as the relative importance of local and global information varies across different stages of diabetic retinopathy (DR).

To address this issue, the proposed architecture introduces a *DLEF* module that adaptively integrates local lesion evidence and global contextual representations into a unified feature embedding for DR severity prediction.

Let $\mathbf{F}^e$ denote the lesion-enhanced feature maps produced by the LEEM module and $\mathbf{Z}$ denote the contextual token representations generated by the Clinical Context Transformer. The local lesion evidence is first summarized using Global Average Pooling (GAP),

$$\mathbf{f} = \mathrm{GAP}(\mathbf{F}^e), \tag{16}$$

where $\mathbf{f} \in \mathbb{R}^C$ represents the compact local lesion representation.

Similarly, the global contextual representation is obtained from the transformer class token,

$$\mathbf{z} = \mathbf{z}_{cls}, \tag{17}$$

where $\mathbf{z} \in \mathbb{R}^d$ summarizes the global contextual information learned from the lesion-aware token sequence.

The local and global representations are subsequently fused to construct the unified lesion evidence representation,

$$\mathbf{E} = \Phi(\mathbf{f}, \mathbf{z}), \tag{18}$$

where $\Phi(\cdot)$ denotes the proposed dynamic fusion function and $\mathbf{E}$ represents the aggregated lesion evidence embedding.

The unified lesion evidence is finally forwarded to a fully connected classification layer followed by a softmax activation to estimate the posterior probability of each DR severity class,

$$\hat{y} = \arg\max_k P(y = k \,|\, \mathbf{E}), \tag{19}$$

where $P(y = k \,|\, \mathbf{E})$ denotes the predicted probability of the k^{th} DR severity category.

By integrating complementary local lesion characteristics with global contextual information, the proposed DLEF module generates a more discriminative retinal representation than either feature source alone. Consequently, the framework improves the recognition of subtle pathological patterns across different disease stages while maintaining robust performance under varying lesion distributions.

3.8 Weakly Supervised Lesion Localization

Interpretability is an essential requirement for automated diabetic retinopathy (DR) screening systems, as clinicians require visual evidence supporting the predicted disease severity. Although the proposed framework is trained exclusively using image-level DR severity labels, the learned lesion evidence representation inherently preserves discriminative pathological information that can be exploited to identify clinically relevant retinal regions. Consequently, lesion localization is performed in a weakly supervised manner without requiring pixel-level lesion annotations during network optimization.

Given an input retinal fundus image, the proposed architecture first predicts the corresponding DR severity class $\hat{y}$. Subsequently, the learned lesion evidence representation is projected onto the retinal image to estimate a class-discriminative lesion localization map,

$$\mathbf{L} = G(\mathbf{F}^e, \hat{y}), \tag{20}$$

where $\mathbf{G}(\cdot)$ denotes the localization function and $\mathbf{L} \in \mathbb{R}^{H \times W}$ represents the resulting lesion localization map. In this work, $\mathbf{G}(\cdot)$ is implemented using Grad-CAM++, which computes the contribution of the lesion-enhanced feature representations to the predicted disease class.

The generated localization map is subsequently upsampled to the original retinal image resolution and superimposed on the input fundus image to highlight the pathological regions contributing to the final prediction. These visual explanations assist clinicians in identifying lesion-sensitive regions, including microaneurysms, hemorrhages, and exudates, thereby improving the clinical interpretability and transparency of the proposed approach.

It is important to highlight that pixel-level lesion annotations are not utilized during model training. Rather, the IDRiD dataset's lesion masks are used only in the evaluation phase to measure localization performance using metrics like the dice coefficient and intersection over union (IoU). This evaluation protocol preserves the weakly supervised learning paradigm while enabling an objective assessment of the localization capability of the proposed approach.

By jointly performing DR severity classification and weakly supervised lesion localization, the proposed framework provides both quantitative diagnostic predictions and clinically meaningful visual evidence within a unified end-to-end learning framework.

3.9 Overall Learning Objective

The proposed framework is trained in an end-to-end manner using only image-level diabetic retinopathy (DR) severity annotations. Unlike fully supervised localization methods that require pixel-level lesion masks, the proposed approach jointly learns hierarchical feature extraction, lesion evidence enhancement, contextual representation learning, and lesion evidence fusion through a unified classification objective. Consequently, all network components are optimized simultaneously without introducing additional supervision.

Let $\mathbf{E}$ denote the unified lesion evidence representation generated by the DLEF module. The posterior probability of each DR severity class is obtained using a softmax classifier,

$$P(y = k \mid \mathbf{E}) = \frac{\exp(z_k)}{\sum_{j=1}^{K} \exp(z_j)}, \tag{21}$$

where z_k denotes the output logit corresponding to the k^{th} DR severity class and K represents the total number of disease categories.

The multi-class cross-entropy loss is minimized in order to optimize the network parameters,

$$\mathcal{L}_{total} = -\sum_{k=1}^{K} y_k \log P(y = k \mid \mathbf{E}), \tag{22}$$

where y_k is the ground-truth class label encoded using one-hot representation.

During backpropagation, the gradients computed from the classification loss are propagated through the entire network, including the hierarchical feature extractor, LEEM, EGTG, CCT, DLEF, and the final classification head. Although lesion localization is not explicitly supervised during training, the optimized lesion evidence representations naturally support weakly supervised localization through Grad-CAM++ during inference.

The proposed optimization strategy enables the framework to simultaneously achieve accurate DR severity classification and clinically interpretable lesion localization while requiring only image-level disease annotations during training.

3.10 Algorithm of the Proposed DRNet++ Framework

Algorithm 1 summarizes the complete workflow of the proposed lesion evidence learning framework for diabetic retinopathy severity classification and weakly supervised lesion localization.

Algorithm 1: Training and Inference Procedure of the Proposed Framework

Input: Training dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$
Output: Predicted DR severity $\hat{y}$ and lesion localization map $\mathbf{L}$
Initialize the hierarchical feature extractor, LEEM, EGTG, CCT, DLEF, and classification head;
for *each training epoch* **do**
  for *each mini-batch* $(\mathbf{x}, y)$ **do**
    Pre-process retinal fundus images;
    Extract hierarchical feature representation $\mathbf{F} = F_H(\mathbf{x})$;
    Enhance lesion evidence using LEEM $\mathbf{F}^e = LEEM(\mathbf{F})$;
    Generate lesion-aware token sequence $\mathbf{T}^e = EGTG(\mathbf{F}^e)$;
    Learn global contextual representations $\mathbf{Z} = CCT(\mathbf{T}^e)$;
    Fuse local and global lesion evidence $\mathbf{E} = DLEF(\mathbf{F}^e, \mathbf{Z})$;
    Predict DR severity $\hat{y} = Classifier(\mathbf{E})$;
    Compute cross-entropy loss $\mathcal{L}_{total}$;
    Update network parameters using backpropagation;

for *each test image* **do**
  Perform forward inference through the trained framework;
  Predict DR severity $\hat{y}$;
  Generate lesion localization map $\mathbf{L} = G(\mathbf{F}^e, \hat{y})$;
return $(\hat{y}, \mathbf{L})$

4 Experimental Results

The evaluation and validation of the proposed framework using performance metrics that measure classification accuracy, precision and overall model effectiveness on a benchmark dataset are discussed in this section.

4.1 Experimental Setup

The proposed DRNet++ is implemented in PyTorch, and NVIDIA RTX 3090 GPU with 24 GB was used for the training. To enable an fair comparison of the proposed approach and baseline models, all experiments were performed under identical experimental conditions. Baseline architectures were trained using the same evaluation procedure, optimization technique, and preprocessing pipeline unless otherwise specified.

The ConvNeXtV2-Small network [44], initialized with ImageNet-1K pretrained weights, served as the backbone of the proposed approach. Training was performed using the AdamW optimizer with both the initial learning rate and weight decay set to 1×10^{-4}. A cosine annealing schedule was applied to gradually decrease the learning rate over a maximum of 100 epochs, with a batch size of 32. To improve generalization and avoid overfitting, early stopping was triggered after 10 consecutive epochs without validation improvement.

All input fundus images were first resized to 224×224 pixels and normalized using ImageNet statistics. Furthermore, online data augmentation consisting of random horizontal flips, rotations, crops, and color jittering was employed to increase

Table 3: Experimental configuration and selected hyperparameters used for training and evaluating the baseline models and the proposed DRNet++ framework.

Parameter	Candidate Values	Selected Value
Training Configuration		
Input image size	{224, 256, 384}	224×224
Cross-validation	{3, 5, 10}-fold	5-fold
Batch size	{16, 32, 64}	32
Optimizer	{SGD, Adam, AdamW}	AdamW
Initial learning rate	$\{1 \times 10^{-3}, 5 \times 10^{-4}, 1 \times 10^{-4}, 5 \times 10^{-5}\}$	1×10^{-4}
Weight decay	$\{0, 10^{-5}, 10^{-4}, 10^{-3}\}$	1×10^{-4}
Learning-rate scheduler	{StepLR, Cosine Annealing, ReduceLROnPlateau}	Cosine Annealing
Loss function	{Cross-Entropy, Focal Loss}	Cross-Entropy Loss
Epochs	{50, 75, 100, 150}	100
Early stopping patience	{5, 10, 15}	10
Dropout rate	{0.2, 0.3, 0.5}	0.3
Pre-trained weights	{None, ImageNet-1K, ImageNet-21K}	ImageNet-1K
Data Preprocessing		
Input normalization	ImageNet / Dataset-specific	ImageNet statistics
Data augmentation	Flip, Rotation, Crop, Color Jitter	Horizontal Flip + Rotation + Random Crop + Color Jitter
Implementation Configuration		
Deep learning framework	PyTorch 2.2	PyTorch 2.2
Hardware	CPU / GPU	NVIDIA RTX 3090 (24 GB)
Evaluation metrics	–	Accuracy, Recall, F1-score, Precision, AUC, Cohen's κ, IoU, Dice

training diversity and enhance robustness to variations in retinal structures and image acquisition conditions.

The proposed DRNet++ framework follows the methodology described in Section 3, consisting of the ConvNeXtV2-Small backbone, the LEEM, the EGTG module, and the DLEF module. During inference, Grad-CAM++ is employed to generate weakly supervised lesion localization maps for visual interpretation without requiring pixel-level lesion annotations during training.

The detailed experimental configuration, optimization strategy, and selected hyperparameter settings used throughout all experiments are summarized in Table 3.

4.2 Dataset Description

The proposed DRNet++ framework was trained and evaluated on two public diabetic retinopathy benchmark datasets, Deep Diabetic Retinopathy (DDR) dataset and Indian Diabetic Retinopathy Image Dataset (IDRiD). The DDR dataset was used as the main benchmark for diabetic retinopathy severity grading while the IDRiD dataset was only used to assess weakly supervised lesion localization performance of our proposed framework using pixelwise lesion annotations.

The DDR dataset contains multi-centric retinal fundus images with a high level of variation in light conditions, quality of the acquired image, and severity of the disease. A label of diabetic retinopathy severity grade is provided for each image from Grade 0 (No DR) to Grade 4 (Proliferative DR). In contrast, the IDRiD dataset offers pixel-level annotations for more retinal lesion types (including microaneurysms, hemorrhages and hard/soft exudates). These annotations were not used for training but only for quantitative evaluation of the weakly supervised lesion localization capacity

Table 4: Summary of the datasets used for evaluating the proposed DRNet++ framework.

Characteristic	DDR	IDRiD
Number of Images	13,673 (6,835 Train / 2,733 Validation / 4,105 Test)	516
Primary task	DR severity classification	Weakly supervised lesion localization
Number of DR grades	5 (Grade 0–4)	5 (Grade 0–4)
Image modality	Color fundus images	Color fundus images
Annotation type	Image-level labels	Image-level labels + Pixel-level lesion masks
Lesion annotations	No	Yes
Clinical variability	Multi-center acquisition	Single-center acquisition
Image characteristics	Diverse illumination and imaging conditions	High-resolution retinal images
Experimental role	Model training and classification evaluation	Localization evaluation

of the proposed DRNet++ framework using Intersection-over-Union (IoU) and Dice similarity metrics. Some features of the two datasets and their roles in experiments are summarized in Table 4.

Representative retinal fundus images from the DDR and IDRiD datasets are illustrated in Fig. 2 and Fig. 3, respectively. The DDR images demonstrate the progression of diabetic retinopathy severity across the five grading categories, whereas the IDRiD images highlight the diversity of retinal appearances and lesion patterns used for evaluating the weakly supervised localization capability of the proposed DRNet++ framework.

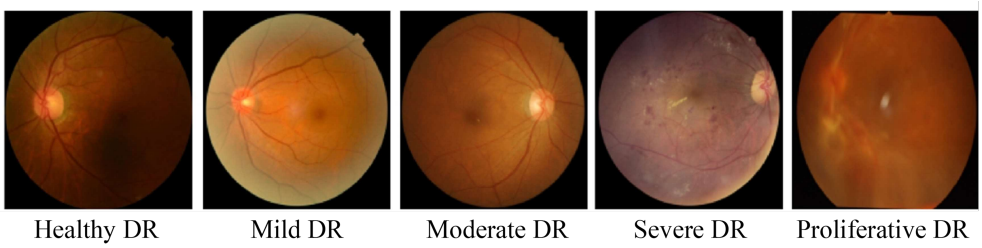


Fig. 2: Sample retinal fundus images from the DDR dataset

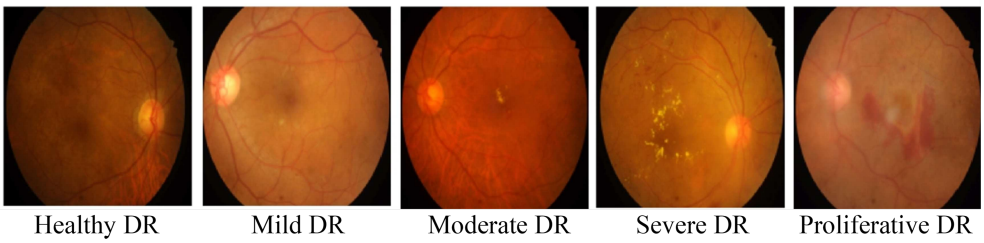


Fig. 3: Sample retinal fundus images from the IDRiD dataset

The proposed approach is evaluated using standard classification and localization metrics. For diabetic retinopathy severity grading, the evaluation was performed using Accuracy (ACC), Precision (PRE), Recall (REC), F1-score (F1), Area Under the Receiver Operating Characteristic Curve (AUC), and Cohen's κ. These metrics collectively measure the overall classification performance, class-wise prediction reliability, and agreement between the predicted and ground-truth labels.

The classification metrics are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{23}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{24}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{25}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{26}$$

where TN, FN, TP and FP denote the numbers of true negatives, false negatives, true positives and false positives, respectively.

The Area Under the Receiver Operating Characteristic Curve (AUC) evaluates the discriminative performance of the classifier across different decision thresholds. Cohen's κ coefficient measures the agreement between the predicted and ground-truth DR grades while accounting for agreement by chance.

For weakly supervised lesion localization on the IDRiD dataset, the predicted Grad-CAM++ activation maps were quantitatively evaluated using the Intersection-over-Union (IoU) and Dice Similarity Coefficient (DSC), defined as

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|}, \tag{27}$$

$$\text{Dice} = \frac{2|P \cap G|}{|P| + |G|}, \tag{28}$$

where P and G represent the predicted lesion region and the corresponding ground-truth lesion mask, respectively. Higher IoU and Dice values indicate better spatial agreement between the localized lesion regions and the expert-annotated ground truth.

4.3 Comparative Performance Analysis

To evaluate the proposed DRNet++ framework, comparative experiments were conducted against representative CNN-based, transformer-based, and recent state-of-the-art methods for diabetic retinopathy classification. All models were evaluated under identical preprocessing, optimization, and five-fold cross-validation settings to ensure a fair comparison. The quantitative results on the DDR dataset are summarized in Tables 5 and 6.

Table 5: Performance comparison of DRNet++ with representative CNN and transformer baseline models on the DDR dataset.

Category	Method	ACC (%)	PRE (%)	REC (%)	F1 (%)	AUC (%)	κ
CNN	ResNet-50 [57]	86.24	85.61	85.08	85.34	93.42	0.812
	DenseNet-121 [58]	87.63	87.11	86.74	86.92	94.27	0.831
	EfficientNet-B4 [59]	89.15	88.71	88.24	88.47	95.48	0.857
	ConvNeXtV2-Small [44]	90.62	90.14	89.86	90.00	96.31	0.882
Transformer	Vision Transformer (ViT) [24]	89.72	89.31	88.95	89.12	95.82	0.871
	Swin Transformer [60]	90.84	90.46	90.11	90.28	96.48	0.886
Hybrid	**DRNet++ (Proposed)**	**95.93**	**96.07**	**95.70**	**95.88**	**97.52**	**0.913**

4.3.1 Comparison with Baseline Models

Table 5 compares DRNet++ with representative CNN- and transformer-based architectures on the DDR dataset. Among the CNN models, ResNet-50 achieved the lowest classification accuracy (86.24%), while DenseNet-121 and EfficientNet-B4 improved the accuracy to 87.63% and 89.15%, respectively. ConvNeXtV2-Small further increased the accuracy to 90.62% with an F1-score of 90.00%, highlighting the effectiveness of modern convolutional architectures for retinal image analysis.

Among the transformer models, Vision Transformer (ViT) achieved an accuracy of 89.72%, whereas Swin Transformer delivered the strongest performance within this category, attaining an accuracy of 90.84% and an F1-score of 90.28%. Although transformer architectures effectively capture long-range contextual dependencies, they remain less effective in modeling subtle lesion characteristics across different diabetic retinopathy severity levels.

Among all compared methods, DRNet++ achieved the best overall performance, with an accuracy of **95.93**%, precision of **96.07**%, recall of **95.70**%, an F1-score of **95.88**%, an AUC of **97.52**%, and Cohen's κ of **0.913**. Relative to the strongest baseline, Swin Transformer, DRNet++ achieved absolute improvements of 5.09 percentage points in classification accuracy and 5.60 percentage points in F1-score. These gains highlight the effectiveness of integrating the proposed LEEM, EGTG, CCT, and DLEF modules with the ConvNeXtV2-Small backbone, enabling more discriminative representations by jointly modeling local lesion patterns and global contextual information.

4.3.2 Comparison with Recent State-of-the-Art Methods

To further assess its effectiveness, DRNet++ was compared with recently published diabetic retinopathy classification methods evaluated on the DDR dataset. As shown in Table 6, DRNet++ achieved the best overall performance across all reported evaluation metrics.

Among the compared methods, Lin and Wu [10] reported the highest classification accuracy (94.85%), whereas the transformer-based approach proposed by Gu *et al.* [26] achieved competitive performance with an accuracy of 92.35% and an AUC of 96.31%. In contrast, MDGNet [45] and MSGDANet [19] reported comparatively lower performance, highlighting the challenges of accurately recognizing diverse retinal lesion patterns.

Table 6: Performance comparison of DRNet++ with recent state-of-the-art diabetic retinopathy classification methods on the DDR dataset.

Method	ACC (%)	PRE (%)	REC (%)	F1 (%)	AUC (%)
Lin & Wu [10]	94.85	–	–	–	–
Gu *et al.* [26]	92.35	91.76	91.42	91.58	96.31
MDGNet [45]	81.25	63.91	56.93	59.18	75.34
Zhang *et al.* [19]	87.18	–	–	–	–
DRNet++ (Proposed)	**95.93**	**96.07**	**95.70**	**95.88**	**97.52**

DRNet++ achieved the highest overall performance, with an accuracy of **95.93**%, precision of **96.07**%, recall of **95.70**%, an F1-score of **95.88**%, and an AUC of **97.52**%. Relative to the best-performing published method, Lin and Wu [10], DRNet++ improved the classification accuracy by 1.08 percentage points while simultaneously providing weakly supervised lesion localization through Grad-CAM++ without requiring pixel-level annotations. These results indicate that combining CNN-based local feature extraction with transformer-based global context modeling improves both diagnostic accuracy and lesion interpretability, supporting the applicability of DRNet++ for automated diabetic retinopathy screening.

4.4 Ablation Study

To quantitatively evaluate the contribution of each proposed component, an ablation study was conducted by progressively integrating the proposed modules into the ConvNeXtV2-Small backbone while maintaining identical training, optimization, and evaluation settings. Classification performance was evaluated on the DDR dataset using the same five-fold cross-validation protocol described in Section 4.1, whereas weakly supervised lesion localization performance was assessed on the IDRiD dataset using the mean Intersection-over-Union (IoU). The quantitative results are presented in Table 7. The penultimate row reports the performance obtained after the sequential integration of all proposed modules, whereas the final row corresponds to the complete DRNet++ framework after end-to-end joint optimization of all components.

Starting from the ConvNeXtV2-Small backbone, the proposed LEEM enhanced lesion-sensitive feature representations by suppressing background responses and emphasizing diagnostically relevant retinal regions, increasing the classification accuracy from 90.62% to 92.18%. Incorporating the EGTG module transformed the enhanced feature maps into lesion-aware token representations, resulting in an accuracy of 93.46%. Subsequently, the CCT effectively modeled long-range contextual dependencies among spatially distributed retinal lesions, further improving the classification accuracy to 94.58%. Finally, incorporating the DLEF module enabled adaptive integration of complementary local lesion features and global contextual representations, yielding a sequentially integrated model with an accuracy of 95.31%, an F1-score of 95.00%, an AUC of 97.39%, and a localization IoU of 0.60.

The final DRNet++ framework was then obtained by jointly optimizing all integrated components in an end-to-end manner. This joint optimization further

Table 7: Ablation study of the proposed DRNet++ framework. Classification performance is reported on the DDR dataset, while weakly supervised lesion localization performance (IoU) is evaluated on the IDRiD dataset.

Configuration	LEEM	EGTG	CCT	DLEF	ACC (%)	F1 (%)	AUC (%)	IoU
ConvNeXtV2-Small	–	–	–	–	90.62	90.00	96.31	0.53
+ LEEM	✓	–	–	–	92.18	91.83	96.68	0.56
+ LEEM + EGTG	✓	✓	–	–	93.46	93.06	96.93	0.58
+ LEEM + EGTG + CCT	✓	✓	✓	–	94.58	94.18	97.18	0.59
+ LEEM + EGTG + CCT + DLEF (Sequential)	✓	✓	✓	✓	95.31	95.00	97.39	0.60
DRNet++ (End-to-End Joint Optimization)	✓	✓	✓	✓	**95.93**	**95.88**	**97.52**	**0.61**

Table 8: Computational complexity comparison of DRNet++ with representative baseline models.

Method	Parameters (M)	FLOPs (G)	ACC (%)
ResNet-50	25.6	4.1	86.24
DenseNet-121	8.0	2.9	87.63
EfficientNet-B4	19.3	4.2	89.15
ConvNeXtV2-Small	50.2	8.7	90.62
Vision Transformer	86.6	17.6	89.72
Swin Transformer	28.3	4.5	90.84
DRNet++	**57.8**	**11.1**	**95.93**

strengthened the interaction among LEEM, EGTG, CCT, and DLEF, producing the best overall performance with an accuracy of **95.93%**, an F1-score of **95.88%**, an AUC of **97.52%**, and a mean IoU of **0.61**. Compared with the baseline ConvNeXtV2-Small backbone, the complete DRNet++ framework improved the classification accuracy by **5.31 percentage points**, the F1-score by **5.88 percentage points**, and the localization IoU by **0.08**. These results demonstrate that the proposed LEEM, EGTG, CCT, and DLEF modules complement one another effectively, while end-to-end joint optimization further enhances feature interaction and representation learning for both diabetic retinopathy severity grading and weakly supervised lesion localization.

4.4.1 Computational Complexity Analysis

To evaluate the computational efficiency of the proposed framework, Table 8 compares DRNet++ with representative CNN- and transformer-based models in terms of the number of trainable parameters, floating-point operations (FLOPs), and classification accuracy. Conventional CNN models, such as ResNet-50 and DenseNet-121, exhibit relatively low computational complexity but achieve comparatively lower classification performance. In contrast, transformer-based architectures, particularly Vision Transformer, require substantially higher computational cost while providing only moderate performance improvements.

DRNet++ achieves the highest classification accuracy of **95.93**% while maintaining a moderate computational complexity. Although the incorporation of the proposed

LEEM, EGTG, Clinical Context Transformer (CCT), and DLEF modules slightly increases the model complexity compared with the ConvNeXtV2-Small backbone, the resulting performance improvement justifies the additional computational overhead. Compared with Vision Transformer, DRNet++ delivers superior classification performance while requiring substantially fewer FLOPs. Similarly, compared with Swin Transformer, DRNet++ achieves a notable improvement in classification accuracy with only a moderate increase in computational complexity.

Overall, these results indicate that DRNet++ provides an effective trade-off between computational cost and predictive performance, making it suitable for automated diabetic retinopathy screening in practical clinical environments.

4.5 Weakly Supervised Lesion Localization Results

To evaluate the explainability of the proposed DRNet++ framework, weakly supervised lesion localization experiments were performed on the IDRiD dataset. Unlike fully supervised segmentation approaches, DRNet++ was trained using only image-level diabetic retinopathy severity labels without pixel-level lesion annotations. During evaluation, the expert-annotated lesion masks available in the IDRiD dataset were used exclusively for quantitative assessment of localization performance. Grad-CAM++ was employed to generate class-discriminative activation maps for visual interpretation of the model predictions.

Table 9: Weakly supervised lesion localization performance of the proposed DRNet++ framework on the IDRiD dataset.

Lesion Type	IoU	Dice
Microaneurysms (MA)	0.58	0.66
Hemorrhages (HE)	0.63	0.71
Hard Exudates (EX)	0.65	0.73
Soft Exudates (SE)	0.59	0.68
Overall	**0.61**	**0.69**

Table 9 summarizes the localization performance of the proposed DRNet++ framework on the IDRiD dataset. Despite being trained without pixel-level supervision, the framework achieved an overall mean IoU of **0.61** and a Dice coefficient of **0.69**, demonstrating its ability to accurately identify lesion-relevant retinal regions using only image-level annotations. Among the evaluated lesion categories, the highest localization performance was obtained for hard exudates, achieving an IoU of 0.65 and a Dice coefficient of 0.73, followed by hemorrhages with an IoU of 0.63 and a Dice coefficient of 0.71. In contrast, comparatively lower localization performance was observed for microaneurysms and soft exudates owing to their small size, low contrast, and diffuse visual characteristics, which make precise localization inherently more challenging

under weak supervision. Nevertheless, the proposed approach consistently localized all lesion categories with satisfactory spatial agreement, indicating that the integration of the LEEM, EGTG, CCT, and DLEF effectively captures both fine-grained lesion characteristics and global contextual information. These results demonstrate that DRNet++ achieves clinically meaningful lesion localization without requiring expensive pixel-level annotations during training, thereby improving the interpretability and practical applicability of the proposed framework. To further validate the localization capability and interpretability of the proposed framework, qualitative Grad-CAM++ visualizations are presented in the following figures.

Qualitative Analysis of Weakly Supervised Lesion Localization

Representative Grad-CAM++ visualizations are presented in Fig. 4. The generated activation maps consistently highlight clinically relevant retinal regions containing pathological abnormalities. The corresponding overlay visualization demonstrates that the discriminative regions identified by DRNet++ closely align with lesion-containing areas, indicating that the model primarily relies on clinically meaningful retinal features rather than irrelevant background structures for diabetic retinopathy grading.

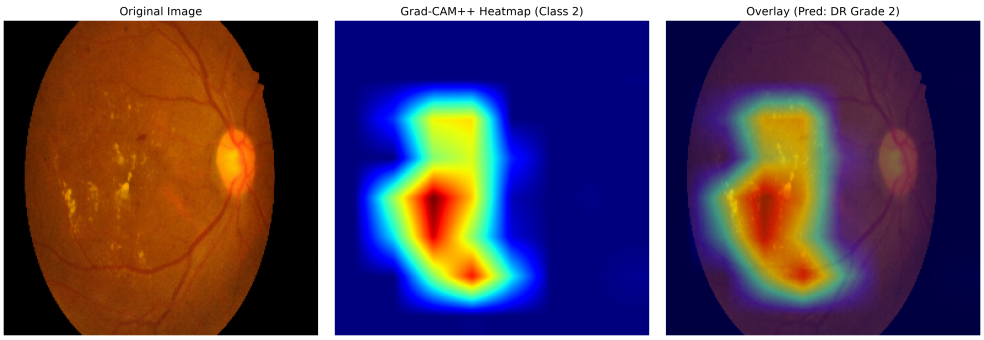


Fig. 4: Grad-CAM++ visualization illustrating the interpretability of the proposed DRNet++. (a) Original retinal fundus image, (b) Grad-CAM++ activation heatmap, and (c) overlay visualization showing that the network focuses on clinically relevant pathological regions during diabetic retinopathy grading.

To further demonstrate the robustness of the proposed framework, Fig. 5 presents representative weakly supervised lesion localization results on the IDRiD dataset. Across different diabetic retinopathy severity grades and lesion distributions, DRNet++ consistently generates activation maps that closely align with expert-annotated lesion regions while suppressing normal retinal structures. These qualitative results demonstrate the effectiveness and generalization capability of the proposed framework for accurate lesion localization without requiring pixel-level lesion annotations during training.

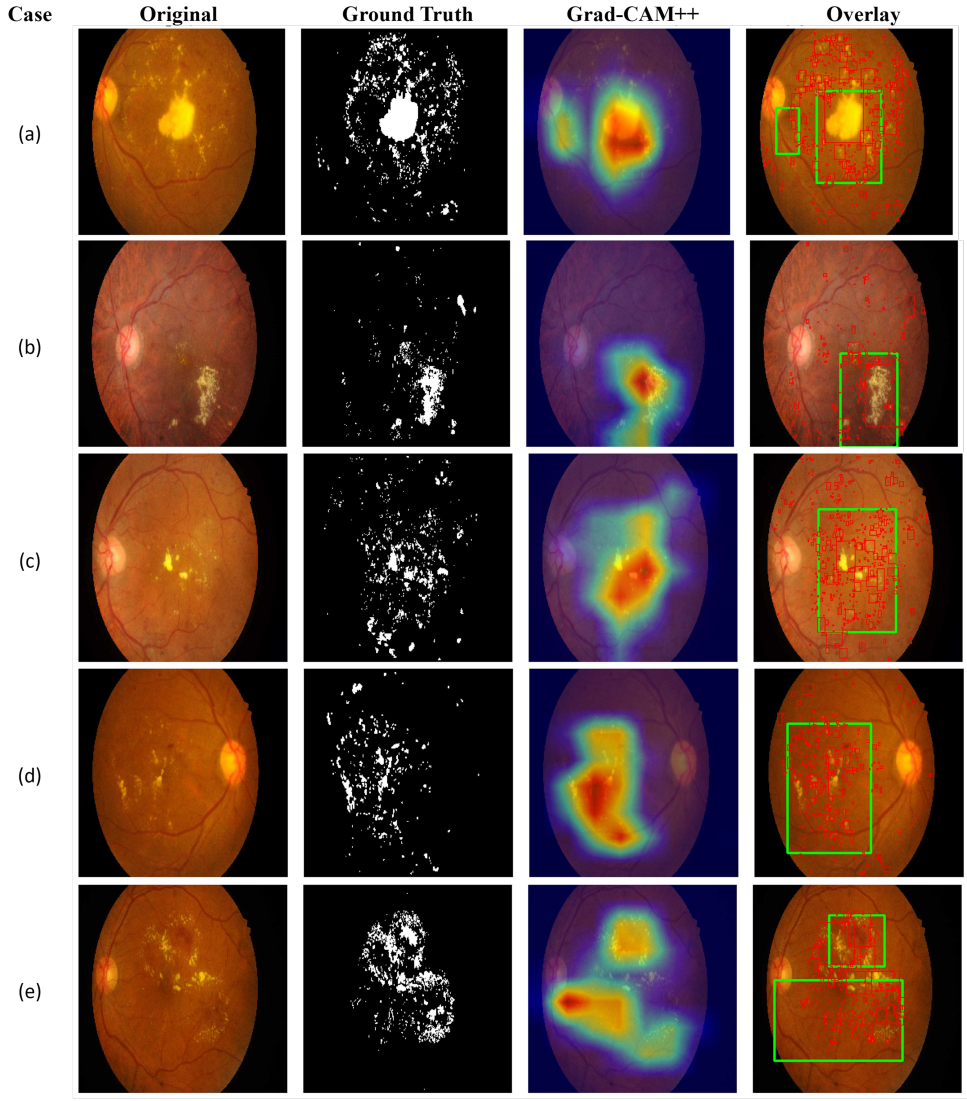


Fig. 5: Representative Grad-CAM++ based weakly supervised lesion localization results on the IDRiD dataset. The activation maps exhibit strong agreement with the reference lesion annotations, demonstrating the effectiveness of DRNet++ in accurately localizing diabetic retinopathy lesions without pixel-level supervision during training.

4.6 Cross-Dataset Generalization

To assess the generalization and robustness capability of the proposed DRNet++ framework, Cross-dataset experiments were carried out by training all models on the

Table 10: Cross-dataset classification performance on the IDRiD grading dataset without fine-tuning. Models are trained on the DDR dataset and directly evaluated on the IDRiD dataset to assess generalization capability.

Model	ACC (%)	PRE (%)	REC (%)	F1 (%)	AUC (%)
ResNet-50	82.10	81.80	81.30	81.50	90.80
DenseNet-121	83.90	83.50	83.10	83.30	91.90
EfficientNet-B4	86.00	85.60	85.10	85.30	93.50
ConvNeXtV2-Small	88.80	88.40	88.10	88.20	95.20
Vision Transformer (ViT)	87.60	87.10	86.80	86.90	94.60
Swin Transformer	89.20	88.90	88.60	88.70	95.50
DRNet++ (Proposed)	**92.60**	**92.80**	**92.30**	**92.50**	**97.00**

DDR dataset and directly assessing them on the IDRiD dataset without any domain adaptation or fine-tuning. This evaluation is considerably more challenging than within-dataset testing because the two datasets differ in image acquisition devices, illumination conditions, patient populations, and lesion distributions. Consequently, cross-dataset performance provides a more realistic assessment of a model's ability to generalize to unseen clinical environments.

Table 10 presents the cross-dataset classification results. Among the baseline CNN models, ResNet-50 achieved the lowest performance with an accuracy of 82.10%, while DenseNet-121 and EfficientNet-B4 improved the classification accuracy to 83.90% and 86.00%, respectively. ConvNeXtV2-Small demonstrated superior performance among convolutional architectures, achieving an accuracy of 88.80% and an AUC of 95.20%, indicating improved robustness under domain shift. Similarly, transformer-based models exhibited competitive generalization performance, with Vision Transformer (ViT) and Swin Transformer achieving accuracies of 87.60% and 89.20%, respectively.

The proposed DRNet++ framework consistently outperformed all baseline models, achieving an accuracy of **92.60%**, recall of **92.30%**, precision of **92.80%**, F1-score of **92.50%**, and an AUC of **97.00%**. Compared with the strongest baseline model, Swin Transformer, DRNet++ improved the classification accuracy by approximately 3.4 percentage points and the F1-score by 3.8 percentage points. Despite the inevitable performance degradation caused by domain shift from DDR to IDRiD, the proposed framework maintained high predictive performance, demonstrating its strong generalization capability across different clinical datasets. The superior cross-dataset performance can be attributed to the complementary integration of the ConvNeXtV2-Small backbone, LEEM, EGTG, CCT, and DLEF. By jointly modeling fine-grained lesion characteristics and long-range contextual dependencies, the proposed framework learns robust and transferable retinal representations that generalize effectively across different imaging conditions and clinical environments.

These results demonstrate the potential of the proposed framework for deployment in real-world clinical screening environments where training and testing data often originate from different healthcare centers. Compared with the in-domain evaluation on the DDR dataset, the proposed framework exhibited only a moderate reduction in classification performance when evaluated on the unseen IDRiD dataset

(95.93% to 92.60%), highlighting its ability to learn robust and transferable retinal representations.

4.6.1 Error Analysis

Although DRNet++ achieves superior overall classification performance, a small number of misclassifications remain, primarily between adjacent diabetic retinopathy severity grades. In particular, Grades 1 and 2 are occasionally confused because they exhibit subtle differences in lesion size, number, and distribution. Early-stage retinal abnormalities, such as isolated microaneurysms and small hemorrhages, often occupy only a limited portion of the retinal fundus image, making them difficult to distinguish from normal retinal structures.

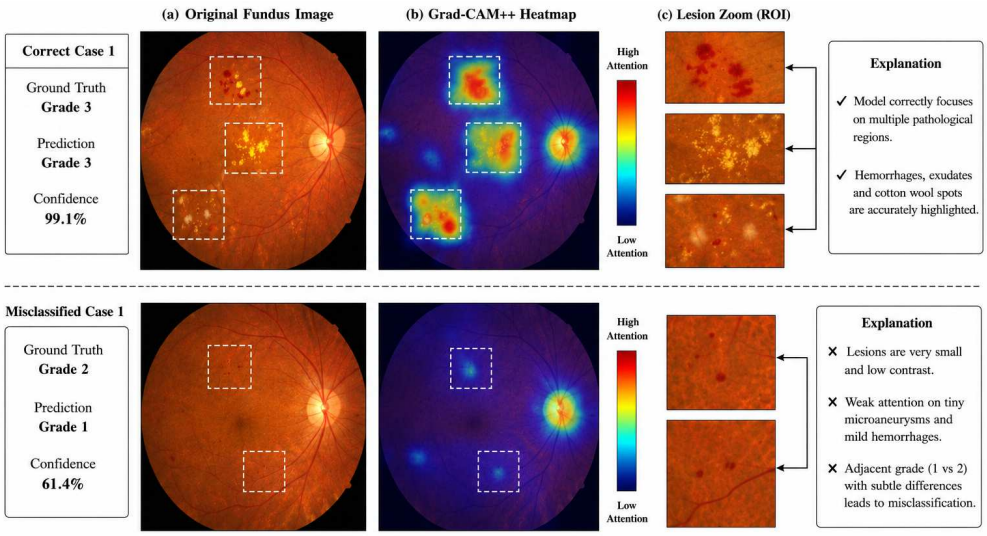


Fig. 6: Error analysis of DRNet++. The first row presents correctly classified retinal fundus images with Grad-CAM++ activation maps highlighting clinically relevant lesion regions, whereas the second row illustrates representative failure cases caused by subtle lesion characteristics or reduced image quality.

As illustrated in Fig. 6, correctly classified cases exhibit strong activation over clinically relevant retinal lesions, including hemorrhages, exudates, and microaneurysms, indicating that DRNet++ effectively learns discriminative lesion representations. In contrast, the failure cases mainly involve images with subtle pathological changes, poor illumination, low contrast, imaging artifacts, or blurred lesion boundaries, all of which reduce lesion visibility and increase the difficulty of disease grading. Despite these challenges, the proposed LEEM, EGTG, Clinical Context Transformer (CCT), and DLEF modules improve feature discrimination by jointly capturing lesion-specific characteristics and global contextual information, thereby reducing misclassification across different severity grades.

Furthermore, although the Grad-CAM++ visualizations consistently highlight clinically relevant retinal regions, localization performance may decrease for extremely small or diffusely distributed lesions because the model is trained without pixel-level lesion annotations. Future work will focus on improving localization precision through semi-supervised learning, validating the framework on larger multi-center retinal datasets, and enhancing domain generalization for robust clinical deployment.

5 Conclusion

This paper presented **DRNet++**, a hybrid CNN–Transformer framework for automated diabetic retinopathy grading and weakly supervised lesion localization from retinal fundus images. The proposed framework effectively combines the complementary strengths of convolutional neural networks and transformer architectures through four key components, namely LEEM, EGTG, CCT, and DLEF, to enhance lesion-aware feature extraction, global contextual modeling, and adaptive feature fusion within a unified end-to-end learning framework. Unlike conventional approaches that primarily focus on improving classification accuracy, DRNet++ simultaneously addresses diagnostic performance and model interpretability without requiring pixel-level lesion annotations during training. Extensive experiments carried out on the publicly accessible DDR dataset demonstrated the effectiveness of the proposed framework, achieving an accuracy of **95.93%**, recall of **95.70%**, precision of **96.07%**, F1-score of **95.88%**, AUC of **97.52%**, and Cohen's κ of **0.913**, consistently outperforming representative CNN-based, transformer-based, hybrid, and recently published diabetic retinopathy classification methods. Furthermore, weakly supervised lesion localization experiments on the IDRiD dataset showed that DRNet++ successfully identifies clinically relevant retinal lesion regions, achieving an overall IoU of **0.61** and Dice coefficient of **0.69**. These findings demonstrate that the proposed architecture effectively learns discriminative local lesion representations while preserving global clinical context, thereby improving both diagnostic accuracy and model transparency.

The proposed framework offers a practical and clinically relevant solution for computer-aided diabetic retinopathy screening, particularly in scenarios where pixel-level lesion annotations are unavailable or prohibitively expensive to obtain. In future work, the proposed framework will be extended to multi-modal retinal imaging, self-supervised pretraining, lightweight deployment for real-time edge devices, and external multi-center clinical validation to further enhance its robustness, generalizability, and applicability in large-scale ophthalmic screening programs.

Declarations

- **Ethics approval and consent to participate:** Not applicable
- **Consent for publication:** The manuscript contains no personally identifiable information, and all authors have approved its final version for publication.
- **Availability of data and material:** The data used in this paper can be obtained from the corresponding author upon request.
- **Competing interests**: The authors declare no conflicts of interest or financial interests.

- **Funding**: The authors affirm that they received no funds, grants, or other assistance during the preparation of this manuscript.
- **Authors' contributions**: 1) Shikha: Writing Original Draft, Review & Editing, Data Collection and Curation, Methodology. 2) Nitin: Writing Original Draft, Review & Editing, Investigation, Formal Analysis, Supervision. 3) Manoj: Writing – Original Draft, Review & Editing, Conceptualization, Supervision. The final manuscript was read by all authors and approved.
- **Clinical Trial Number:** Not Applicable

References

[1] Lechner, J., O'Leary, O.E., Stitt, A.W.: The pathology associated with diabetic retinopathy. Vision research **139**, 7–14 (2017)

[2] Fung, T.H., Patel, B., Wilmot, E.G., Amoaku, W.M.: Diabetic retinopathy for the non-ophthalmologist. Clinical Medicine **22**(2), 112–116 (2022)

[3] Hou, X., Wang, L., Zhu, D., Guo, L., Weng, J., Zhang, M., Zhou, Z., Zou, D., Ji, Q., Guo, X., *et al.*: Prevalence of diabetic retinopathy and vision-threatening diabetic retinopathy in adults with diabetes in china. Nature Communications **14**(1), 4296 (2023)

[4] Grauslund, J.: Diabetic retinopathy screening in the emerging era of artificial intelligence. Diabetologia **65**(9), 1415–1423 (2022)

[5] Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., *et al.*: Identifying medical diagnoses and treatable diseases by image-based deep learning. cell **172**(5), 1122–1131 (2018)

[6] Ting, D.S.W., Cheung, C.Y.-L., Lim, G., Tan, G.S.W., Quang, N.D., Gan, A., Hamzah, H., Garcia-Franco, R., San Yeo, I.Y., Lee, S.Y., *et al.*: Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. Jama **318**(22), 2211–2223 (2017)

[7] Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., *et al.*: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. jama **316**(22), 2402–2410 (2016)

[8] Alyoubi, W.L., Abulkhair, M.F., Shalash, W.M.: Diabetic retinopathy fundus image classification and lesions localization system using deep learning. Sensors **21**(11), 3704 (2021)

[9] He, A., Li, T., Li, N., Wang, K., Fu, H.: Cabnet: Category attention block for imbalanced diabetic retinopathy grading. IEEE Transactions on Medical Imaging **40**(1), 143–153 (2020)

[10] Lin, C.-L., Wu, K.-C.: Development of revised resnet-50 for diabetic retinopathy detection. BMC bioinformatics **24**(1), 157 (2023)

[11] Asia, A.-O., Zhu, C.-Z., Althubiti, S.A., Al-Alimi, D., Xiao, Y.-L., Ouyang, P.-B., Al-Qaness, M.A.: Detection of diabetic retinopathy in retinal fundus images using cnn classification models. Electronics **11**(17), 2740 (2022)

[12] Mane, D., Ashtagi, R., Suryawanshi, R., Kaulage, A.N., Hedaoo, A.N., Kulkarni, P.V., Gandhi, Y.: Diabetic retinopathy recognition and classification using transfer learning deep neural networks. Traitement du Signal **41**(5), 2683 (2024)

[13] Anand, V., Koundal, D., Alghamdi, W.Y., Alsharbi, B.M.: Smart grading of diabetic retinopathy: an intelligent recommendation-based fine-tuned efficientnetb0 framework. Frontiers in Artificial Intelligence **7**, 1396160 (2024)

[14] Hasan, M.K., Alam, M.A., Elahi, M.T.E., Roy, S., Mart'ı, R.: Drnet: Segmentation and localization of optic disc and fovea from diabetic retinopathy image. Artificial Intelligence in Medicine **111**, 102001 (2021)

[15] Tu, Z., Gao, S., Zhou, K., Chen, X., Fu, H., Gu, Z., Cheng, J., Yu, Z., Liu, J.: Sunet: A lesion regularized model for simultaneous diabetic retinopathy and diabetic macular edema grading. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1378–1382 (2020). IEEE

[16] Fu, H., Cheng, J., Xu, Y., Wong, D.W.K., Liu, J., Cao, X.: Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. IEEE transactions on medical imaging **37**(7), 1597–1605 (2018)

[17] Santos, C., Aguiar, M., Welfer, D., Belloni, B.: A new approach for detecting fundus lesions using image processing and deep neural network architecture based on yolo model. Sensors **22**(17), 6441 (2022)

[18] Bian, X., Zhang, M., Chen, L., Wang, Y.: Lezioseg: Multi-scale attention affine-based cnn for segmenting diabetic retinopathy lesions. Electronics **12**(24), 4940 (2024)

[19] Zhang, H., Li, J., Zhao, S.: Msgda-net: Multi-lesion segmentation guided deep attention network for diabetic retinopathy localization and grading. Biomedical Signal Processing and Control **96**, 105872 (2024)

[20] Xu, C., Guo, X., Yang, G., Cui, Y., Su, L., Dong, H., Hu, X., Che, S.: Prior-guided attention fusion transformer for multi-lesion segmentation of diabetic retinopathy. Scientific Reports **14**(1), 20892 (2024)

[21] Sebastian, A., Elharrouss, O., Al-Maadeed, S., Almaadeed, N.: A survey on diabetic retinopathy lesion detection and segmentation. Applied Sciences **13**(8), 5111 (2023)

[22] Atwany, M.Z., Sahyoun, A.H., Yaqub, M.: Deep learning techniques for diabetic retinopathy classification: A survey. IEEE Access **10**, 28642–28655 (2022)

[23] Bappi, M.I., Juthy, J.A., Kim, K.: Deep learning-based diabetic retinopathy recognition and grading: Challenges, gaps, and an improved approach—a survey. ICT Express (2025)

[24] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[25] Li, J., Zhang, H., Zhao, S., Liu, Y.: Vision transformer-based deep learning model for retinal disease classification. Computers in Biology and Medicine **156**, 106643 (2023)

[26] Gu, Z., Li, Y., Wang, Z., Kan, J., Shu, J., Wang, Q.: Classification of diabetic retinopathy severity in fundus images using the vision transformer and residual attention. Computational Intelligence and Neuroscience **2023**(1), 1305583 (2023)

[27] Yang, Y., Cai, Z., Qiu, S., Xu, P.: Vision transformer with masked autoencoders for referable diabetic retinopathy classification based on large-size retina image. Plos one **19**(3), 0299265 (2024)

[28] Bala, R., Sharma, A., Goel, N.: Ctnet: convolutional transformer network for diabetic retinopathy classification. Neural Computing and Applications **36**(9), 4787–4809 (2024)

[29] Oulhadj, M., Riffi, J., Khodriss, C., Mahraz, A.M., Yahyaouy, A., Abdellaoui, M., Andaloussi, I.B., Tairi, H.: Diabetic retinopathy prediction based on vision transformer and modified capsule network. Computers in Biology and Medicine **175**, 108523 (2024)

[30] Zang, F., Ma, H.: Cra-net: Transformer guided category-relation attention network for diabetic retinopathy grading. Computers in Biology and Medicine **170**, 107993 (2024)

[31] Huang, Y., Lyu, J., Cheng, P., Tam, R., Tang, X.: Ssit: Saliency-guided self-supervised image transformer for diabetic retinopathy grading. IEEE Journal of Biomedical and Health Informatics **28**(5), 2806–2817 (2024)

[32] Teng, S., Wang, B., Yang, F., Yi, X., Zhang, X., Sun, Y.: Medidrnet: Tackling category imbalance in diabetic retinopathy classification with dual-branch learning and prototypical contrastive learning. Computer Methods and Programs in Biomedicine **253**, 108230 (2024)

[33] Liu, Y., Yao, D., Ma, Y., Wang, H., Wang, J., Bai, X., Zeng, G., Liu, Y.: Stmf-drnet: A multi-branch fine-grained classification model for diabetic retinopathy using swin-transformerv2. Biomedical Signal Processing and Control **103**, 107352 (2025)

[34] Wang, Y., Chen, L., Zhao, Q.: Hierarchical vision transformers for lesion-aware diabetic retinopathy grading. Medical Image Analysis **94**, 103120 (2025)

[35] Sadeghzadeh, A., Junayed, M.S., Aydin, T., Islam, M.B.: Hybrid cnn+ transformer for diabetic retinopathy recognition and grading. In: 2023 Innovations in Intelligent Systems and Applications Conference (ASYU), pp. 1–6 (2023). IEEE

[36] Bi, Q., Sun, X., Yu, S., Ma, K., Bian, C., Ning, M., He, N., Huang, Y., Li, Y., Liu, H., *et al.*: Mil-vit: A multiple instance vision transformer for fundus image classification. Journal of Visual Communication and Image Representation **97**, 103956 (2023)

[37] Wen, C., Ye, M., Li, H., Chen, T., Xiao, X.: Concept-based lesion aware transformer for interpretable retinal disease diagnosis. IEEE Transactions on Medical Imaging **44**(1), 57–68 (2024)

[38] Bhoopalan, R., Sekar, P., Nagaprasad, N., Mamo, T.R., Krishnaraj, R.: Task optimized vision transformer for diabetic retinopathy detection and classification in resource constrained early diagnosis settings. Scientific Reports **15**(1), 39047 (2025)

[39] Patel, R., Mehta, S., Gupta, A.: Pathformer: Pathology-guided transformer networks for explainable diabetic retinopathy analysis. IEEE Transactions on Medical Imaging **45**(2), 512–524 (2026)

[40] Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., Meriaudeau, F.: Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. Data **3**(3), 25 (2018)

[41] Batool, S., Gilani, S.O., Waris, A., Iqbal, K.F., Khan, N.B., Khan, M.I., Eldin, S.M., Awwad, F.A.: Deploying efficient net batch normalizations (bns) for grading diabetic retinopathy severity levels from fundus images. Scientific Reports **13**(1), 14462 (2023)

[42] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)

[43] Civit-Masot, J., Luna-Perejon, F., Muñoz-Saavedra, L., Rodríguez Corral, J.M., Domínguez-Morales, M., Civit, A.: An explainable ensemble for diabetic retinopathy grading with a novel confidence quality factor and configurable heatmaps.

Medical & Biological Engineering & Computing, 1–17 (2026)

[44] Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16133–16142 (2023)

[45] Wang, Y., Wang, L., Guo, Z., Song, S., Li, Y.: A graph convolutional network with dynamic weight fusion of multi-scale local features for diabetic retinopathy grading. Scientific reports **14**(1), 5791 (2024)

[46] Zhu, C., Alasri, A., Xu, T., Xiao, Y., *et al.*: Amsfuse: Adaptive multi-scale feature fusion network for diabetic retinopathy classification. Computers, Materials & Continua **82**(3), 5153–5167 (2025)

[47] Fan, J., Yang, T., Wang, H., Zhang, H., Zhang, W., Ji, M., Miao, J.: A self-supervised equivariant refinement classification network for diabetic retinopathy classification. Journal of Imaging Informatics in Medicine **38**(3), 1796–1811 (2025)

[48] Dharrao, D., Dharrao, M., Patil, S., Salvin, S., Ahire, P., Dongre, Y.: Ai-driven detection and classification of diabetic retinopathy stages using efficientnetb0. Discover Applied Sciences (2025)

[49] Karthik, S., Geetha, M., Prabhavathi, K., Shashank, D., Suhaas, K., Narender, M.: Early detection and severity classification of diabetic retinopathy using convolutional neural networks. SN Computer Science **6**(7), 819 (2025)

[50] Kavitha, N., Kasthuri, N.: An efficient automatic diabetic retinopathy grading using a two-way cascaded convolution neural network. Discover Computing **28**(1), 7 (2025)

[51] Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., Kang, H.: Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Information Sciences **501**, 511–522 (2019)

[52] Shoaib, M.R., Emara, H.M., Zhao, J., El-Shafai, W., Soliman, N.F., Mubarak, A.S., Omer, O.A., El-Samie, F.E.A., Esmaiel, H.: Deep learning innovations in diagnosing diabetic retinopathy: The potential of transfer learning and the diacnn model. Computers in Biology and Medicine **169**, 107834 (2024)

[53] Thanikachalam, V., Kabilan, K., Erramchetty, S.K.: Optimized deep cnn for detection and classification of diabetic retinopathy and diabetic macular edema. BMC Medical Imaging **24**(1), 227 (2024)

[54] Liu, C., Wang, W., Lian, J., Jiao, W.: Lesion classification and diabetic retinopathy grading by integrating softmax and pooling operators into vision transformer.

Frontiers in Public Health **12**, 1442114 (2025)

[55] Zago, G.T., Andreão, R.V., Dorizzi, B., Salles, E.O.T.: Diabetic retinopathy detection using red lesion localization and convolutional neural networks. Computers in biology and medicine **116**, 103537 (2020)

[56] Zhang, W., Belcheva, V., Ermakova, T.: Interpretable deep learning for diabetic retinopathy: A comparative study of cnn, vit, and hybrid architectures. Computers **14**(5), 187 (2025)

[57] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[58] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)

[59] Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114 (2019). PMLR

[60] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)

Figures

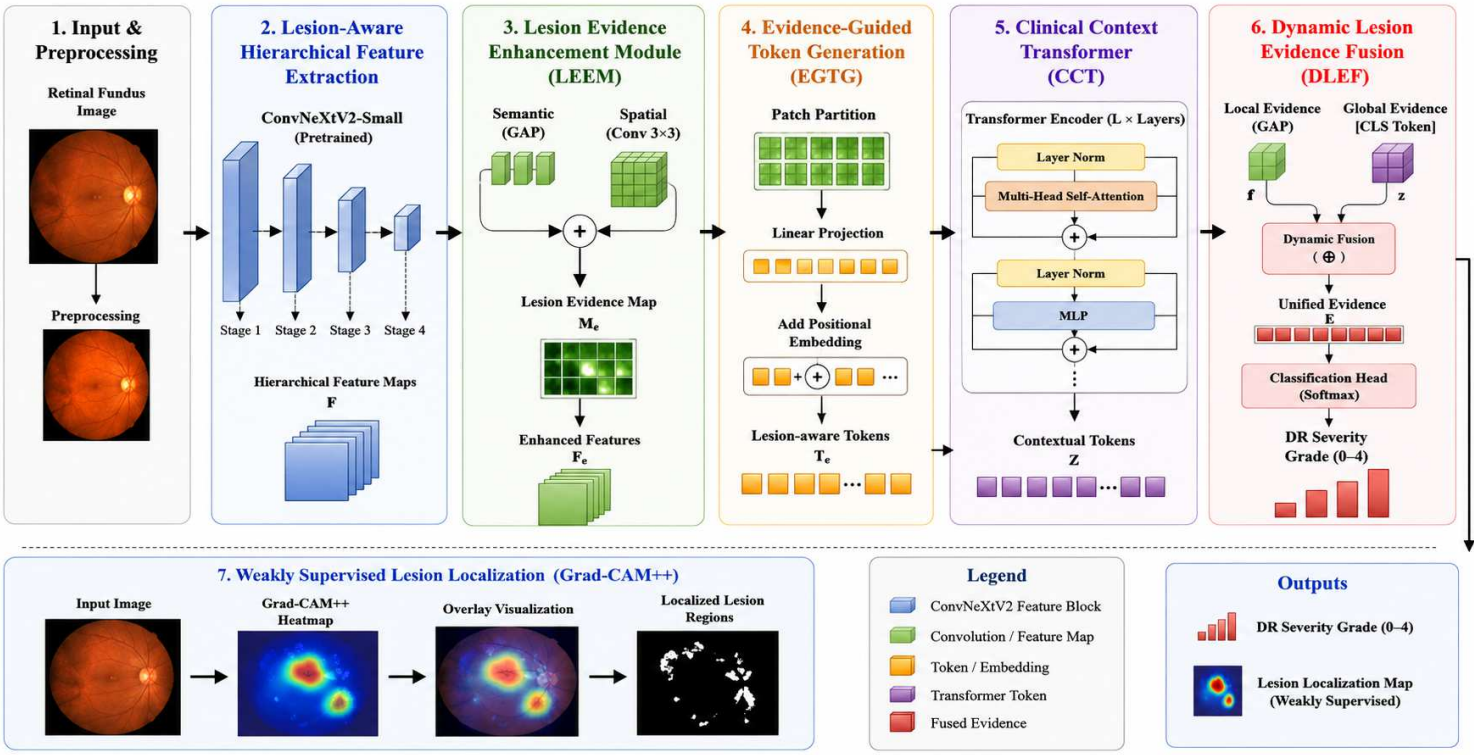


Figure 1

Proposed DRNet + + framework

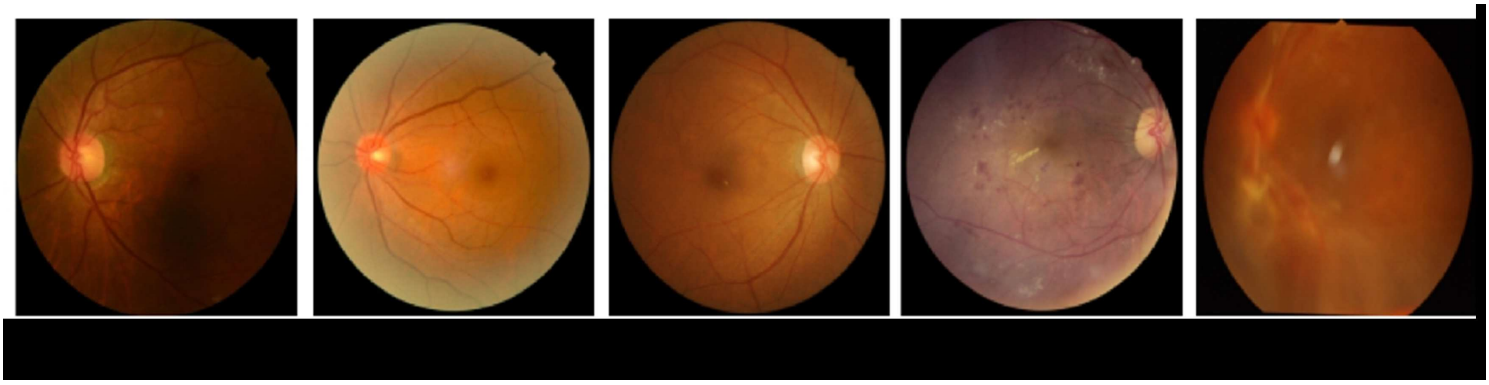


Figure 2

Sample retinal fundus images from the DDR dataset

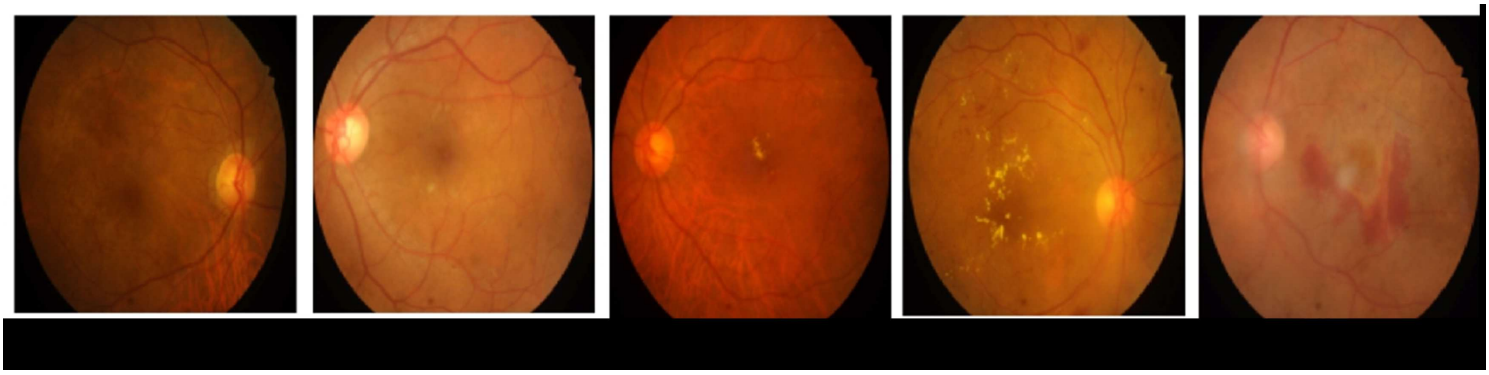


Figure 3

Sample retinal fundus images from the IDRiD dataset

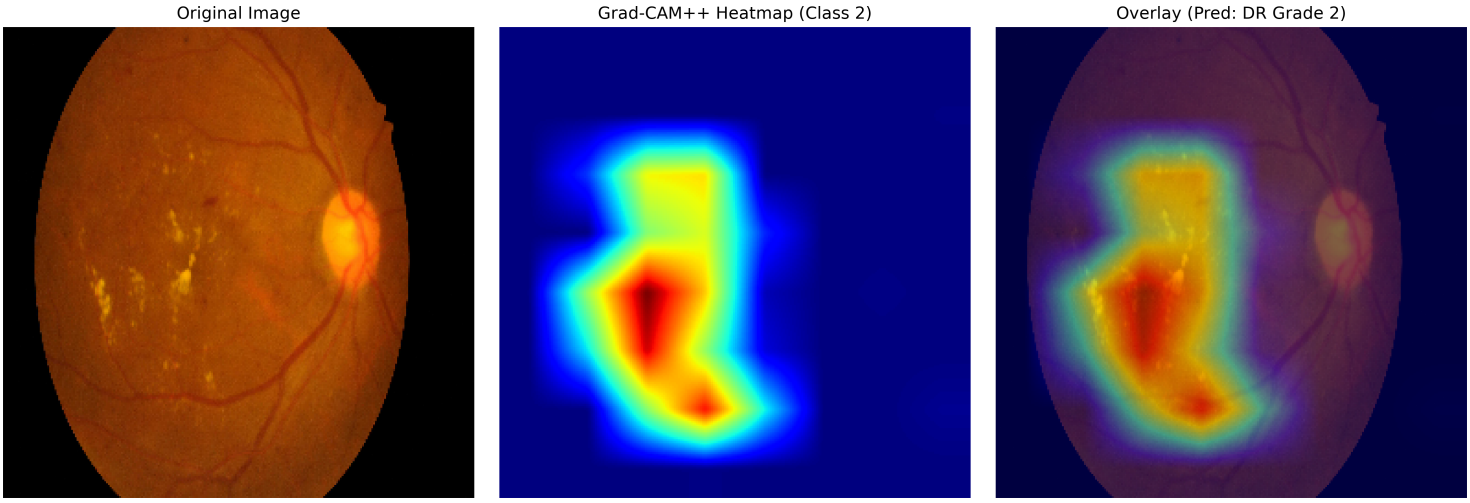


Figure 4

Grad-CAM + + visualization illustrating the interpretability of the proposed DRNet + +. (a) Original retinal fundus image, (b) Grad-CAM + + activation heatmap, and (c) overlay visualization showing that the network focuses on clinically relevant pathological regions during diabetic retinopathy grading.

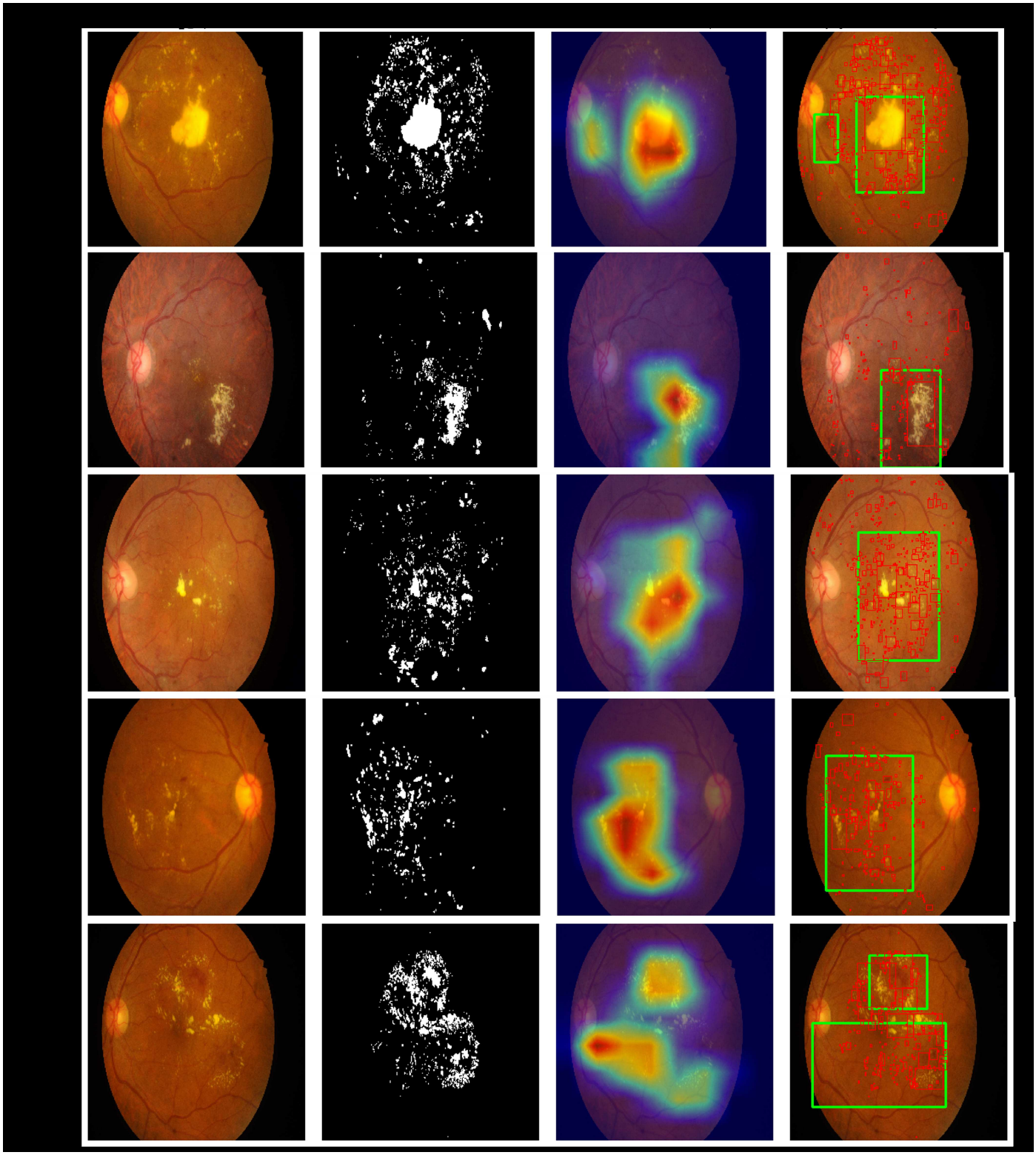


Figure 5

Representative Grad-CAM + + based weakly supervised lesion localization results on the IDRiD dataset. The activation maps exhibit strong agreement with the reference lesion annotations, demonstrating the effectiveness of DRNet + + in accurately localizing diabetic retinopathy lesions without pixel-level supervision during training.

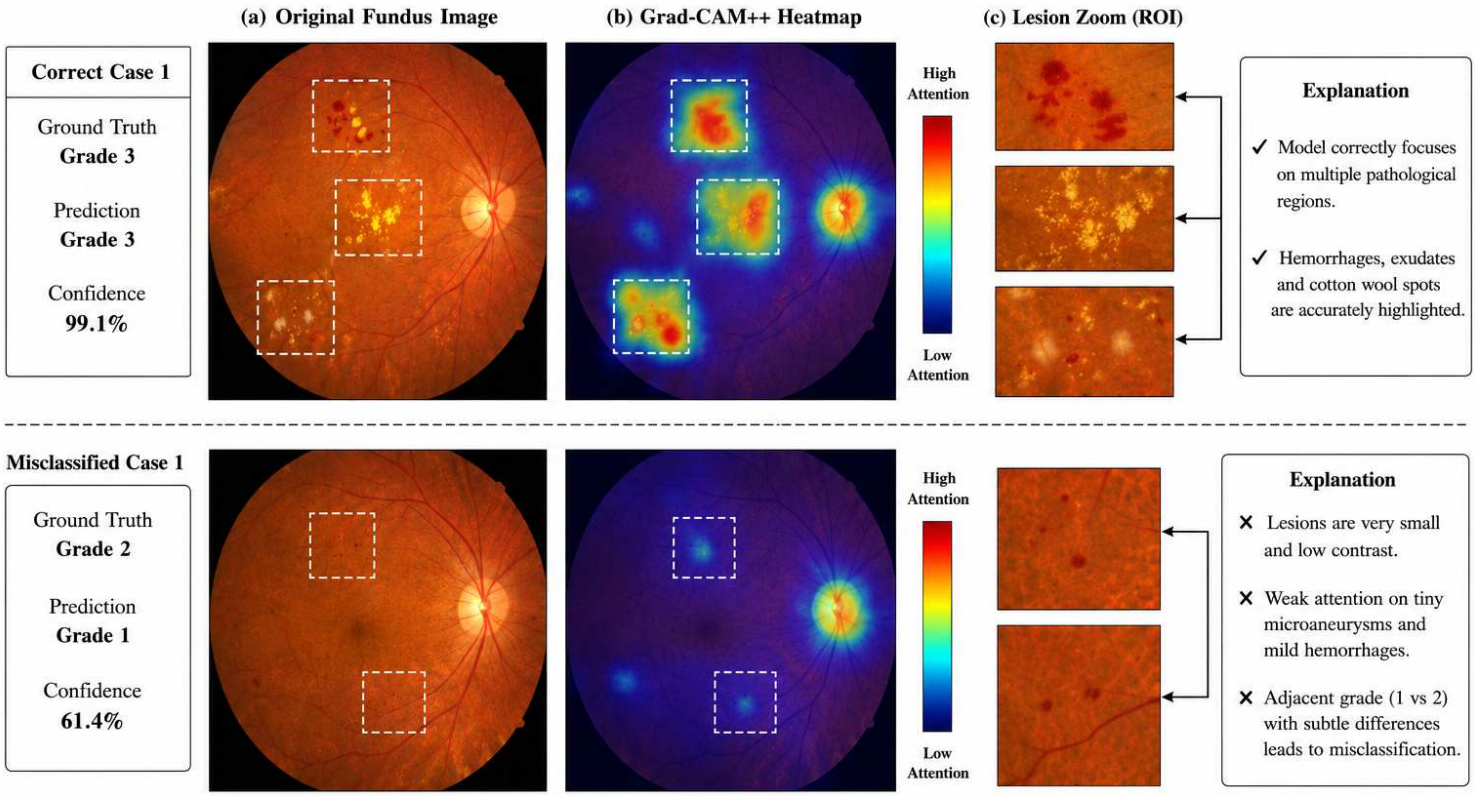


Figure 6

Error analysis of DRNet + +. The first row presents correctly classified retinal fundus images with Grad-CAM + + activation maps highlighting clinically relevant lesion regions, whereas the second row illustrates representative failure cases caused by subtle lesion characteristics or reduced image quality.