An Explainable Multi-Scale Attention-Based Deep Learning Framework for Trustworthy Diabetic Retinopathy Detection Using Enhanced Fundus Images

Muhammad Nabeel Mehmood[1], Muhammad Hassaan Ashraf[1,2*]
[1]Faculty of Computing, Riphah International University, Islamabad, 45200, Pakistan
[2]Department of Computer Science, COMSATS University Islamabad, 45550, Pakistan
E-mail: mhassaan.scholar@gmail.com

Keywords: Artificial intelligence, computer-aided diagnosis, medical images, diabetic retinopathy, convolutional neural network, explainable ai, hybrid CNN, multi scale features, multi-level features, swish activation function, residual connection blocks, inception blocks, explainable deep learning, grad-CAM, channel-wise CLAHE

Received: February 20, 2026

Diabetic Retinopathy (DR) is a microvascular complication caused by prolonged hyperglycaemia that damages the retinal capillary network, putting people with diabetes at risk. Damage to these tiny vessels increases vascular permeability and results in microaneurysms, hemorrhages, and hard exudates. If left untreated, progression may lead to irreversible vision loss. Early detection and timely treatment substantially reduce the risk of blindness. Current clinical diagnosis relies on ophthalmologist assessment, which results in diagnostic delays and inter-observer variability. Inter-class similarity and intra-class variability in Fundus Images (FIs), due to heterogeneity in lesion morphology, retinal pigmentation, and imaging protocols, yields diagnostic ambiguity. Conventional Deep Learning (DL)-based frameworks act as black boxes, yielding high predictive accuracy but lacking interpretable evidence of the image features about their decisions, thereby limiting clinician trust and clinical adoption. To mitigate these limitations, we propose an explainable DL-based mechanism that classifies FIs into different DR severity grades. This study presents a heterogeneous feature-extraction method inspired by residual and inception networks, combined with attention mechanism to learn disease specific features, so that the model focuses on the most relevant patterns. Residual skip connections and multi-scale inception style branches, in combination with features refinement, helps classifying lesion severity with high precision. The proposed framework begins with segmentation of FIs to extract Regions of Interests (ROIs), followed by Channel-wise Contrast Limited Adaptive Histogram Equalization (C-CLAHE) combined with Difference of Gaussian (DoG) filtering, optimizing the color channels independently and standardizing contrast to highlight fine-grained retinal structures. Geometric minority oversampling is employed to ensure balanced representation of all severity categories during training. Proposed framework incorporates Swish activation to enhance gradient propagation and employs the class-aware focal loss formulation to mitigate class imbalance issue, emphasizing hard examples. Experimental results on the IDRiD and APTOS2019 combined test-set demonstrate that the proposed model achieves screening (healthy/diseased) and lesion-severity (mild/moderate/severe/proliferative) classification accuracies of 99.27% and 75.90%, respectively. Although, discriminating the four DR stages pose challenges like severe imbalance and inter-class differences among classes, which are better aided by proposed framework. On the DDR benchmark, the model achieves comprehensive DR-grading accuracy of 80.76% on 5-class classification task. The Heterogeneous Attention Residual-Inception Convolutional Network (HARIC-Net) outperforms several state-of-the-art (SOTA) DL-based architectures, including AlexNet, ResNet, DenseNet and Inception Net, along with recent domain specific hybrids in DR-grading task. HARIC-Net's parameter counts of ~2 million underscores its computational efficiency and robust generalization capability. Gradient-weighted Class Activation Mapping (Grad-CAM), an Explainable Artificial Intelligence (XAI) technique, is used to visualize class-discriminative regions, and to provide interpretable predictions.

Povzetek: Študija predstavi razložljiv HARIC-Net za razvrščanje diabetične retinopatije, ki z ROI segmentacijo, C-CLAHE/DoG predobdelavo, rezidualno-inception značilkami, pozornostjo in Grad-CAM razlagami izboljša zanesljivo ter klinično bolj zaupanja vredno oceno resnosti iz fundusnih slik.

1 Introduction

Diabetic Retinopathy (DR) is a common eye complication of diabetes and a significant cause of visual loss and blindness. DR primarily results from prolonged abnormalities in glucose metabolism, that damages ocular tissues which include retinal blood vessels, nerves, the lens, and the vitreous body [1]. In the initial stages, DR lacks discriminative symptoms. However, as the severity

progresses, the DR patients face visual problems that include floaters, blurred vision, and distortion, which worsen in later stages. DR symptoms include microaneurysms (small red dots), retinal hemorrhages (dark-red spots caused by ruptured vessels), hard exudates (yellow lipid deposits), soft exudates or cotton-wool spots (signs of localized ischemia), and, in advanced proliferative stages, pathological retinal neovascularization (abnormal new vessel growth) [2], as depicted in Figure 1. Consequently, DR can be categorized as no DR, mild DR, moderate DR, severe DR, and proliferative DR.

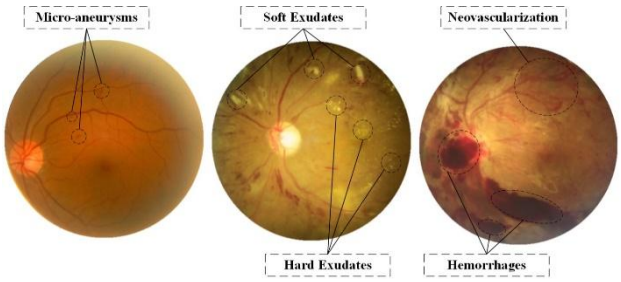

Figure 1: DR lesions highlighted in retinal FIs

Global projections highlight the growing severity of this issue. According to the World Health Organization, diabetes has spread from approximately 200 million individuals in 1990 to about 830 million in 2022 [3], emphasizing the necessity of early DR screening and severity grading, to reduce the burden on healthcare systems. Manual interpretation of high-resolution FIs by ophthalmologists remains the clinical standard, which is resource-intensive and time-consuming. DR lesions manifest diversely, requiring ophthalmologists to manually annotate lesion types, sizes, distributions, and quantities. These procedures rely on the expertise of ophthalmologists and are affected by the limited availability of medical resources. Moreover, inter-class similarity and intra-class variability [4], [5], caused by the heterogeneity of lesion morphology and variations in retinal pigmentation, complicates accurate interpretation and classification. This emphasizes the need for automated Computer-Aided Diagnosis (CAD) systems to support DR screening and improve retinal care [4], [6], [7], [8], [9], [10]. CAD systems integrate feature extraction and classification algorithms that identify changes between normal and affected FIs to construct a feature space, which helps in determining the grade of DR. CAD frameworks have proven valuable across multiple domains, such as intelligent transportation systems, agriculture, and healthcare applications, where robust feature engineering accurate, scalable, and cost-effective diagnostic support systems [11], [12], [13].

Historically, CAD pipelines for DR have progressed as two major paradigms. The classical computer vision approach relies on hand-crafted feature extraction followed by conventional machine-learning classifiers such as Random Forest (RF), Support Vector Machine (SVM), and Naïve Bayes [14], [15]. While these methods achieved reasonable performance, their dependence on domain-specific feature engineering limits generalization across diverse datasets due to variations in imaging devices, illumination conditions, and patient demographics [16]. The advent of high-performance Graphics Processing Units (GPUs) has driven a paradigm shift towards DL. Within DL, Convolutional Neural Networks (CNNs) have become the reliable framework for automated FI analysis, as they learn hierarchical and task-specific feature representations directly from data [17]. More recently, Vision Transformer (ViT) architectures have emerged, but their reliance on large-scale pretraining datasets and substantial computational resources poses challenges for deployment in medical imaging scenarios. Therefore, this study centres on CNN-based framework, emphasizing multi-level and multi-scale feature representations to enhance diagnostic accuracy and model generalization.

Canonical CNN architectures, like VGG, Inception Net, DenseNet, and ResNet, offer valuable design principles for DR classification. VGG demonstrates the effectiveness of using small, fixed-size filters (3×3) stacked deep in the network to capture fine-grained features [18]. Inception excels at capturing multi-scale patterns by employing parallel filters of varying sizes, allowing for more diverse features extraction [19]. DenseNet enhances feature reuse through dense connectivity, facilitating the flow of information across layers [20], and ResNet overcomes the vanishing gradient problem through the use of residual skip connections [21]. Despite these advancements, real-world DR classification faces significant challenges, including varying lesion scales, pronounced class imbalances, and the lack of computationally efficient models suitable for clinical deployment. Furthermore, existing systems struggle to address variations at different feature levels, limiting the full utilization of low-level image features like edges and small spots. These features are often indicative of growing lesion representations, but can be misclassified as noise or other retinal structures. As disease progress, certain early-stage lesions symptoms get overlooked in the decision-making process, further complicating accurate staging. To capture these morphological features, hybrid CNNs are proposed that achieve high accuracies, but are computationally expensive [22]. Moreover, existing CNN-based frameworks act as opaque "black boxes" that lack interpretability [23], [24], [25]. The decision-making process remains non-transparent, limiting its broader clinical adoption. Addressing these issues requires a careful integration of multi-scale and multi-level feature extraction techniques, regularization strategies, interpretability enhancements, and robust preprocessing methods.

To mitigate these challenges, we propose HARIC-Net, a robust framework for automated DR detection and severity grading. Feature extraction plays a critical role in accurate disease classification within framework. Our model integrates residually dense connections with inception-style multiscale parallel filters, which enable it to effectively capture features at varying spatial scales. This approach enhances the network’s ability to refine representations allowing it to capture global and fine-

grained details, though it is prone to background noise. An attention mechanism is introduced to address this issue, that emphasize on clinically relevant retinal regions, while minimizing the influence of background noise, allowing the model to concentrate on the most crucial regions, thereby improving overall diagnostic accuracy, while having very less trainable parameters. The preprocessing pipeline applies CLAHE to each color channel, combined with DoG filtering, to enhance low-level image features. This preprocessing improves the visibility of lesions and retinal vasculature that become blurred due to progressive lesion development. We utilized a Global Average Pooling (GAP) layer for robust feature aggregation while constraining the model's complexity. The Swish activation function is adopted for efficient nonlinear transformations. The class imbalance issue is addressed using class-aware geometric minority oversampling and focal loss function, enhancing model's generalization across varying imaging conditions. For post-hoc interpretability, we generate Grad-CAM saliency maps to emphasize on the regions that contribute to decision-making during FIs classifications. HARIC-Net is evaluated on three publicly available datasets that include IDRiD, APTOS2019, and DDR, encompassing 2-class screening, 4-class lesion grading, and 5-class comprehensive classification tasks. Classification of lesions in DR has difficulties such as imbalance between classes and overlapping patterns. With the predominance of a healthy class in 5-class classification, the model's performance in distinguishing lesion severity can be suppressed, which is why 4-class grading serves as a clinically relevant parameter to accurately assess lesions categorization ability of proposed framework. This work proposes:

- A novel heterogeneous CNN architecture, HARIC-Net, that combines residual connectivity blocks and inception-style convolutions to capture multi-level and multi-scale retinal pathology. Attention blocks are utilized, following each layer in HARIC-Net, to focus on relevant lesion areas. Semantic features aggregation is achieved through GAP head, and the flow of gradients is enhanced with Swish activations, leading to better training stability. HARIC-Net achieved robust accuracy while having only ~2 million (M) trainable parameters.
- A novel preprocessing pipeline that begins with RoI segmentation followed by a channel-wise CLAHE enhancement combined with DoG filtering. This pipeline preserves color fidelity while emphasizing fine vascular and lesion details across individual color channels and improves the visibility of low-contrast and small-scale pathological features in retinal images.
- Integration of class-aware geometric minority oversampling augmentation technique, along with the class-aware focal loss formulation, to reduce disparities in classes. The technique focuses learning on hard examples, improving the strength of model to differentiate various severity grades.
- Application of Grad-CAM as an XAI technique for explaining the decisions of the model, which leads to improved interpretability and enhances generalization to non-homogenous imaging conditions.
- Comprehensive study on IDRiD, APTOS2019, and DDR benchmarks, through comparisons between SOTA baselines, hybrid CNNs, and an ablation study quantifying the contributions of dataset variability, augmentation technique and attention blocks.

The rest of the paper will be structured in the following manner. Section 2 summarizes the related work literature and identifies challenges in automated DR analysis. Section 3 explains the datasets, the preprocessing pipeline, and explains the architecture of the HARIC-Net, and experimental protocols. Section 4 gives the ablation study and the performance evaluation of the proposed model. Section 5 is about comparative results with emphasis on the clinical implications of the findings. Lastly, section 6 concludes the paper and points out the future research directions.

2 Literature review

Classifying DR and severity grading is essential to help ophthalmologists to prescribe timely treatments. In recent years, DL neural network-based frameworks have emerged as promising solutions in various domains [11], [13], [26], particularly for the early diagnosis and prevention from blindness in DR cases [22], [25], [27], [28], [29]. Following review presents how these systems have emerged from basic transfer learning baselines to domain specific hybrids, and contributed to the field, defining the best ways to grade DR.

Venkatasubbu et al. [30] propose a transfer-learning ensemble approach for DR diagnosis, which combines signal-processing enhancements with fine-tuned ImageNet-based backbones. It utilized swarm-sparse decomposition, adaptive histogram equalization, morphological filtering, and time-frequency transforms, to enhance microlesions prior to feature extraction from CNN backbones, including AlexNet, VGG16, ResNet50, and GoogLeNet. Feature Maps (FMs) are standardized, and ensembled for robust predictions. The Local Interpretable Model-Agnostic Explanations (LIME) is employed to highlight important regions involved in decision-making. Design choices are validated through ablation studies on the APTOS2019 dataset. However, the trade-offs include a high computational cost because of numerous preprocessing steps, which explains why lighter alternatives are needed.

Sidiq S and Benil T [31] also proposed a transfer-learning ensemble, which splits a five-level DR grading task into ten one-versus-one (OVO) binary subproblems, making it suitable to be deployed in a low-resource environment. Backbones such as MobileNet, InceptionV3, and DenseNet121 are fine-tuned as per task at hand. The model is trained by the Adam optimizer on augmented 224 × 224 FIs. The calculations of the probabilities obtained through averaging of the results and

predictions based on majority voting were also made to reach the final decision with the help of a calibrated threshold. The use of OVO decomposition was observed through the assessments on APTOS2019, IDRiD, Messidor-2, and DDR. Such an approach compromises the simplicity of making unilateral pair decisions, and it involves the complexity of multi-binary model management, requiring need for simpler architectures.

Suganthi M and Arun M [32] utilized dual-stream representations and domain-aware preprocessing to preserve edge and vessel cues. They integrated wavelet-Retinex denoising with multi-resolution curvelet decompositions, routing towards parallel CNN branches. Moreover, they introduced polynomial frequency-domain activations combined with a constrained Salp-Swarm optimizer to minimize histogram distortion and preserve high-frequency lesion details. Ashwini and Dash [33] proposed a two-stream architecture that separates the vessel pathway (CLAHE applied on green-channel) from the texture pathway (LBP), to fuse mid-level representations to enhance mild DR symptoms. Both works highlight that frequency-aware preprocessing and domain-guided stream fusion can significantly improve early-stage detection.

Ashraf et al.'s DRD-Net [25] adopts a hybrid backbone fusion approach to extract multi-scale and multi-level features for DR detection. DRD-Net integrates VGG16-style depth, DenseNet's dense connectivity, and Inception modules to balance feature propagation with multi-scale RoI extraction. CLAHE is used as a preprocessing technique, while targeted geometric augmentations and an Adam-based training pipeline is employed on a combined IDRiD and APTOS2019 dataset. Soomro et al. [34] fuse mid-level descriptors from fine-tuned ResNet-50 and EfficientNet-B0 using 1×1 standardization and concatenation, followed by downstream classification using XGBoost, SVM, or RF. To address class imbalance, they incorporate extensive augmentations and Synthetic Minority Over-sampling Technique (SMOTE). Both works attribute performance gains to the complementarity of the backbones and multi-level feature aggregation, but lack interpretability.

Madarapu et al. [35] enhance multi-scale feature fusion by incorporating attention mechanisms to improve the detection of subtle lesions and reduce inter-class ambiguity. They propose MuR-CAN, which concatenates dual backbones, Xception and ResNet50v2, and applies sub-pixel mean subtraction and CLAHE. The dimension-reduced features are then processed through a novel Multi-Dilation Attention Block (MDAB), which utilizes depthwise convolutions with varying dilation rates, SE-style channel reweighting, and a combined GAP and max fusion. Flattened attention features are subsequently classified using SVM. Singh et al. [36] fuse low- and high-level backbone FMs and refine them using a multi-scale residual attention block and a cross-attention block, which jointly model inter-channel interactions and spatial dependencies. Both studies demonstrated ablation studies and Grad-CAM analysis to conclude that attention and multi-dilation strategies successfully emphasize lesion-relevant regions.

Ashraf et al.'s [22] Hierarchical-Inception-Residual-Dense Network (HIRD-Net) integrates hierarchical fusion, attention mechanisms, and interpretability into a compact explainable CNN framework. The pre-processing pipeline combines CLAHE with dilated DoG filtering. It incorporates Squeeze-and-Excitation (SE) channel attention modules placed before 4 GAP heads. Hard-Swish is employed with focal loss for improved classification performance. Grad-CAM heatmaps are generated to provide class-discriminative visualizations, enhancing the model's interpretability. Evaluated across IDRiD, APTOS2019, DDR, and EyePACS benchmarks, their proposed architecture effectively preserves microaneurysm-scale details while aggregating broader lesion context, but remains computationally expensive with ~4.8 M parameters.

Abraham S and Kovoor B [37] introduced HAR-Net, that employed integrated saliency learning and explainability during training. A lightweight ResNet-18 encoder was trained, incorporating a Channel-Spatial Attention Encoding (CSAE) block and a Hierarchical Cross-Attention (HCA) refinement stage. The model optimizes these heads using focal loss to mitigate class imbalance. Model's performance was evaluated using explainability metrics on APTOS2019 and external test sets from DDR and IDRiD. Renu D and Saji k [38] presented Attention-AlexNet, which combines green-channel segmentation via FPCOM with a CBAM-augmented 9-layer AlexNet. They tuned hyperparameters using Improved Nutcracker Optimizer, and validated preprocessing, attention, and optimizer variants using non-parametric significance tests.

Ashraf M and Alghamdi H [27] proposed a lightweight Hierarchical Feature-Fusion Network (HFF-Net), that preserve early, multiscale low-level cues, with only ~1.18M parameters. Their pre-processing pipeline extracts the fundus RoIs followed by CLAHE enhancement, and employs selective geometric augmentation. Proposed framework incorporates a four-branch HFF stem that generates multiscale outputs, which are then processed by Scale-Adaptive Dense Connection (SADC) blocks, consisting of pooled, dense (3×3, 5×5), and dilated branches. Each block is followed by GAP head, and after concatenation, the descriptors are fed into a Softmax classifier. Swish activations, batch normalization, and sparse categorical cross-entropy loss stabilize training. Ablation studies shows that HFF stem, CLAHE with targeted augmentation, and the two-block SADC stack, is an optimal design choice for both screening and grading tasks. The HFF-Net framework lacks interpretability, and still have room for improvement in prediction performances.

Across the literature, challenges such as inconsistent handling of class imbalance, limited cross-dataset validation, sparse lesion-level annotations, lack of interpretability, and a need for robust computationally efficient model persist. These issues motivate our study, which focuses on integrating principled preprocessing, efficient feature fusion, and interpretability to advance clinically viable DR diagnosis, with minimal computational cost.

3 Proposed DR diagnosis framework

In the early stages, DR lesions are less symptomatic and may cause only mild vision problems. But in the later stages, it can cause permanent vision loss. As the disease evolves at various stages, diagnosing it early remains a challenging yet important step in preventing blindness. This study proposes a novel framework, HARIC-Net, designed to facilitate early DR diagnosis and lesion severity grading. Following sections provide a detailed analysis of proposed framework's components, including the experimental datasets, preprocessing pipeline, augmentation technique, and explanation of multi-level and multi-scale feature extraction along with attention blocks and XAI-based interpretability. Figure 2 illustrates the primary workflow of the proposed framework.

3.1 Experimental datasets

In this work, we conducted experiments on three publicly available DR datasets to validate the performance of our proposed model, which include the Indian Diabetic Retinopathy Image Dataset (IDRiD) [39], Asia Pacific Tele-Ophthalmology Society 2019 (APTOS2019) [40], and Diagnosis of Diabetic Retinopathy (DDR) [41] datasets. The IDRiD dataset originally consists of 516 color FIs, while APTOS2019 contains 3,663 images, both including healthy (NoDR) and unhealthy (DR) retinas. These datasets were divided into training and testing subsets, with 80% of the samples for training and the remaining 20% for testing. To assess the model's robustness against domain shift challenges, we combined the two datasets, resulting in a diverse dataset of 4,179 images. Table 1 presents the class-wise sample distribution of benchmark datasets utilized in this study. In this context, Class1, Class2, Class3, and Class4 corresponds to mild, moderate, severe, and proliferative DR lesion types.

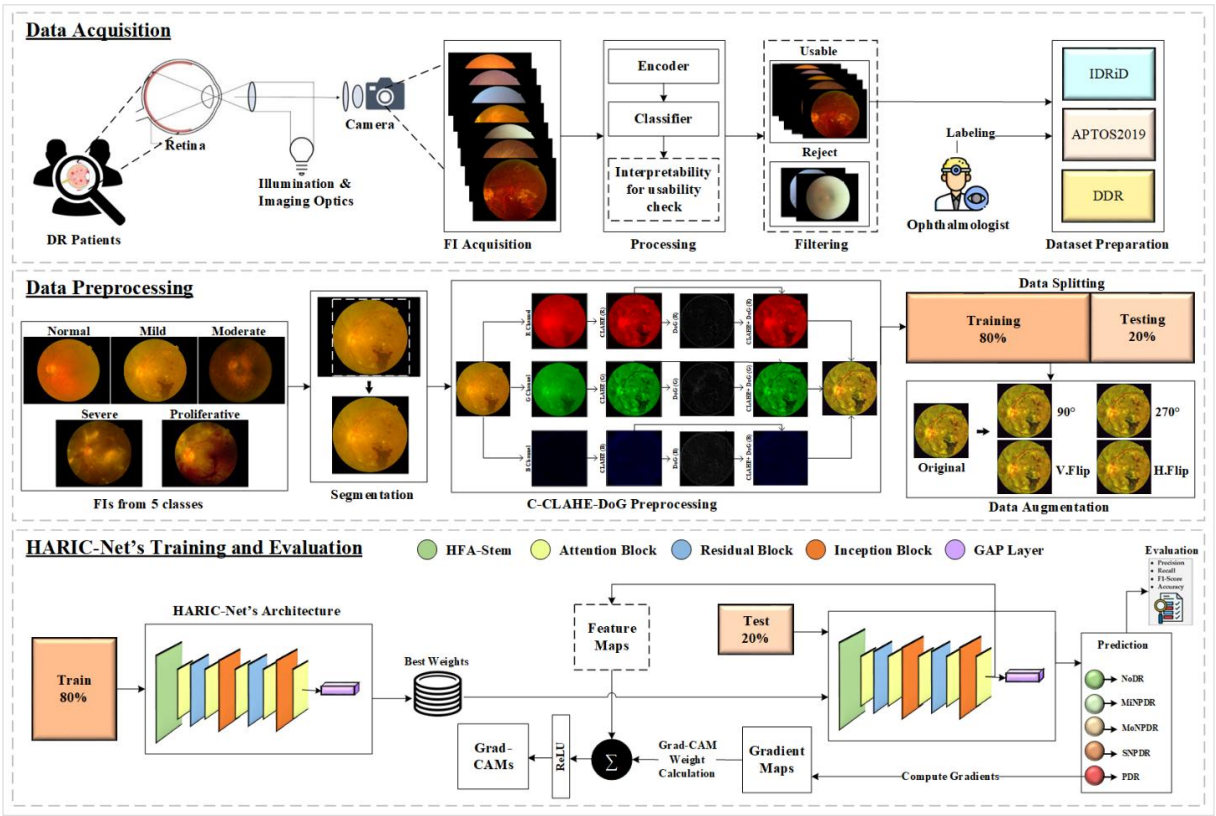

Figure 2: Proposed framework for DR diagnosis

Table 1: Lesion class wise samples for IDRiD, APTOS2019 and DDR benchmarks

Classes	IDRiD		Total samples in each class	APTOS2019		Total samples in each class	DDR		Total samples in each class
	Test (20%)	Train (80%)		Test (20%)	Train (80%)		Test (20%)	Train (80%)	
Class1	5	20	25	74	296	370	126	504	630
Class2	34	134	168	200	799	999	897	3,580	4477
Class3	19	74	93	39	154	193	48	188	236
Class4	13	49	62	59	236	295	183	730	913
Normal	34	134	168	361	1445	1806	1,254	5,012	6266

The DDR dataset is a large collection of Chinese retinal FIs released by the Ocular Disease Intelligent Recognition (ODIR-2019) group. It consists of 13,673 color FIs collected from 147 hospitals across 23 provinces in China. In the original DDR dataset, ophthalmologists annotated the images into six classes, Healthy (NoDR), Mild Non-Proliferative DR (MiNPDR), Moderate Non-Proliferative DR (MoNPDR), Severe Non-Proliferative DR (SNPDR), Proliferative DR (PDR), and an "ungradable" class for poor-quality images. Since the ungradable class contains images of insufficient quality, it was excluded, and a five-class subset of the DDR dataset, comprising 12,522 images, was used for the experiments. This large-scale dataset allows to assess HARIC-Net's robustness and generalization across diverse imaging conditions and patient populations. Figure 3 illustrates sample images of each severity level from all three benchmark datasets.

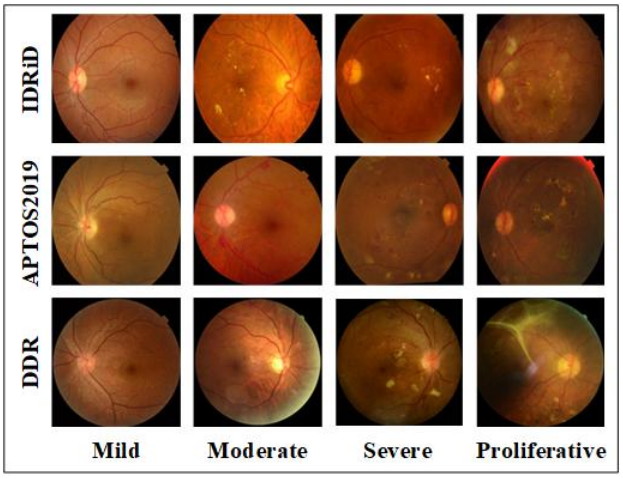

Figure 3: Lesion class samples from each benchmark.

3.2 Data segmentation and preprocessing

Preprocessing pipeline begins by utilizing a straightforward threshold-based technique to extract the retinal RoIs, removing the surrounding blank boundaries. First, each color FI $I(x, y)$ is transformed into grayscale $g(x, y)$. The foreground (retina) and background (near-black border) are then separated using intensity thresholding to create a binary mask $B(x, y)$, that is given as:

$$B(x,y) = \begin{cases} 1, & g(x,y) > T, \\ 0, & g(x,y) \leq T, \end{cases} \quad (1)$$

where T denotes the threshold value, set to $T = 10$. This step enhances retinal visibility while suppressing background artifacts. We then detect the contour to identify connected boundaries. Contour with the largest area is selected as the retinal region, given the retina's dominance within the field of view. A tight bounding rectangle is computed around the selected contour, and the original RGB image is cropped to this RoI. The cropped images are then saved for subsequent processing. Representative examples of the original and cropped images are shown in Figure 4(a) and 4(b), respectively, along with pseudo code of mechanism in Algorithm 1.

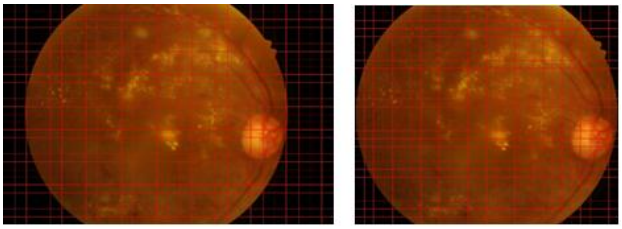

Figure 4: (a) Original Image (4288 × 2848), (b) Cropped Image (3430 × 2848).

Algorithm 1: Pseudo code for extracting RoI from retinal FIs.

```
function ROI-Extraction(OriginalDataset, DestFolder, KernelSize = 5×5, Threshold = t, MinContourArea = a_min)
Create DestFolder if it does not exist // Ensure destination folder exists
ROIDataset = empty set // Initialize the dataset to store ROIs

for each image I in OriginalDataset do
    W, H = width(I), height(I) // Get the image dimensions
    I_gray(p) = 0.299·R(p) + 0.587·G(p) + 0.114·B(p) for every pixel p // Convert to grayscale
    I_blur = GaussianBlur(I_gray, KernelSize) // Apply Gaussian blur to smooth the image
    I_bin(p) = 255 if I_blur(p) > Threshold else 0 for every pixel p // Apply thresholding for binary mask
    I_bin = MorphologicalOpenClose(I_bin) // Optionally clean noise in binary mask
    ContourSet = FindExternalContours(I_bin) // Find external contours in the binary mask
    if ContourSet is empty then
        CONTINUE // Skip if no contours are found
    end if
    C_max = contour in ContourSet with maximum area // Select the largest contour
    if Area(C_max) < MinContourArea then
        CONTINUE // Skip if the contour area is too small
    end if
    (x, y, w, h) = BoundingRect(C_max) // Get bounding box for the largest contour
    if (x == 0 AND y == 0 AND x + w == W AND y + h == H) OR TouchesBorder(C_max, W, H) then
        CONTINUE // Skip if the contour touches the border or is too large
    end if
    ROI = Crop(I, x, y, w, h) // Crop the region of interest (ROI) from the original image
```

```
[24] SaveImage(ROI, DestFolder, basename(I)) // Save the ROI to the destination folder
[25] Add ROI to ROIDataset // Add the ROI to the dataset
[26] end for
[27] return ROIDataset // Return the dataset of ROIs
[28] end function
```

Following retinal cropping and resizing, CLAHE enhancement combined with DoG filtering is applied to each channel individually to enhance lesion and vessel visibility while preserving chromatic information. First, the input color FI is split into three-color channels I_R, I_G, I_B. For each channel $c \in \{R, G, B\}$ CLAHE is applied independently. Instead of applying contrast equalization to the entire image, CLAHE divide it into discrete data sections called tiles, which are used to enhance contrast locally. Each channel is partitioned into an 8 × 8 grid of non-overlapping tiles, and for each tile at location (i, j), the intensity histogram $h_{c,i,j}(k)$ for gray level $k \in \{0, \dots, N-1\}$ is computed. The tile histogram is clipped at a Clip Limit (CL) to limit noise amplification. The clipped histogram bins are redistributed uniformly across all bins. The normalized Cumulative Distribution Function (CDF) for tile (i, j) in channel c is given by:

$$F_{c,i,j}(n) = \frac{1}{M} \cdot \sum_{k=0}^{n} h_{c,i,j}(k); \quad n = 0, \dots, N-1, \quad (2)$$

where M is the number of pixels in the tile and N is the number of gray levels. The histogram of each tile is clipped at $CL = 4.0$ before the CDF computation to limit noise amplification, and any excess counts are redistributed uniformly across the histogram bins.

Adjusting the histogram of each color channel independently emphasize the distinct spectral information that each channel provides about retinal anatomy and pathology. In FIs, the green channel typically offers the highest contrast between blood vessels and background, making it particularly sensitive to early microvascular lesions, such as microaneurysms, blot hemorrhages, and small exudates. In contrast, the red channel better preserves signals from choroidal and deeper retinal structures, which are relevant lesions varying in advanced stages like neovascularization [42]. The blue channel is significantly attenuated by macular pigment and light scatter, leading to fewer diagnostically useful vascular details. However, it may be useful for highlighting superficial lipid deposits in certain cases. Thus, the remapped intensity for a pixel with an original value p_{old} within tile (i, j) is obtained by the standard histogram-equalization mapping:

$$p_{map,c,i,j} = \lfloor (N-1) F_{c,i,j}(p_{old}) \rfloor \quad (3)$$

To prevent block artifacts at the boundaries of tiles, the final output intensity for a pixel located between four neighboring tiles is computed using bilinear interpolation of the corresponding tile mappings. If the pixel's relative distances to the top-left tile along the horizontal and vertical axes are denoted by α and β (where $0 \le \alpha, \beta \le 1$), the interpolated output for channel c is given by:

$$\begin{aligned} & p_{new,c} \\ &= (1-\alpha)(1-\beta)\, p_{map,c,i,j} \\ &+ \quad \alpha(1-\beta)\, p_{map,c,i+1,j} \\ &+ (1-\alpha)\beta p_{map,c,i,j+1}\ \alpha\beta p_{map,c,i+1,j+1} \end{aligned} \quad (4)$$

Applying CLAHE independently to the individual color channels, I_R, I_G, I_B produce locally contrast-enhanced channels, I'_R, I'_G, I'_B, valuable in utilizing the distinct diagnostic contributions of each color band while avoiding potential color artifacts that may arise when applying CLAHE on original RGB image.

Following C-CLAHE, a DoG filter is applied to the enhanced channels, to further emphasize high-frequency lesion structures. For each enhanced channel I'_c, we compute two Gaussian-smoothed versions using kernels with standard deviations σ_1 and σ_2(with $\sigma_2 > \sigma_1$). The 2D Gaussian kernel is defined by:

$$G_\sigma(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right), \quad (5)$$

The DoG for I'_c is then computed as:

$$\begin{aligned} \mathrm{DoG}_c(x, y) = & \left(G_{\sigma_1} * I'_c\right)(x, y) \\ & - \left(G_{\sigma_2} * I'_c\right)(x, y), \end{aligned} \quad (6)$$

where * denotes 2D convolution. We used $\sigma_1 = 1.0$ and $\sigma_2 = 2.0$, corresponding approximately to 5 × 5 and 11 × 11 support kernels, respectively. The DoG filtering mitigates background illumination while subtracting the gaussian-blurred images, highlighting vessel edges and tiny lesion structures such as blot hemorrhages and microaneurysms. Subsequently, high-frequency lesion cues within each channel are reinforced by combining the DoG mask with the corresponding I'_c via additive fusion. The fused channels are then stacked to reconstruct the final enhanced color image:

$$\begin{aligned} & I^{enh}(x, y) \\ &= \left[I'_{R,fused}(x, y), I'_{G,fused}(x, y), I'_{B,fused}(x, y)\right] \end{aligned} \quad (7)$$

Applying DoG to individual I'_c highlights fine vascular and lesion details, as depicted in Figure 5.

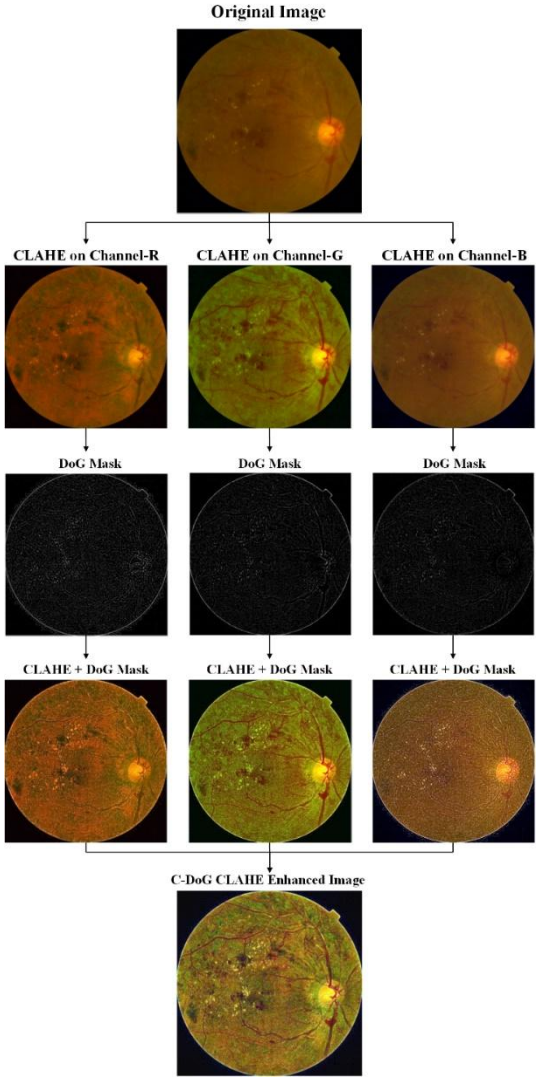

Figure 5: Visual representation of per channel output after applying C-CLAHE-DoG

The C-CLAHE-DoG preprocessing pipeline concentrates on morphological and chromatic lesion features while preserving fine vascular details. Applying CLAHE to the red channel enhanced contrast, highlighting large hemorrhages and abnormal blood vessels. When combined with the DoG mask, it further sharpened their boundaries. The green channel provides the strongest contrast for vessels and microaneurysms, and applying CLAHE-DoG fusion particularly produced rigid vessel outlines and enhanced the visibility of blot lesions, that are critical for early-stage detection. The blue channel, though noisier, highlights small, bright deposits such as hard exudates and surface irregularities. Consequently, per-channel pre-processing and additive fusion enhance high-frequency lesion signatures, while subsequent clipping and normalization prevent color clipping, thereby preserving a natural RGB balance in the reconstructed FIs. Representative outputs of the C-CLAHE-DoG preprocessing pipeline are shown in Figure 6.

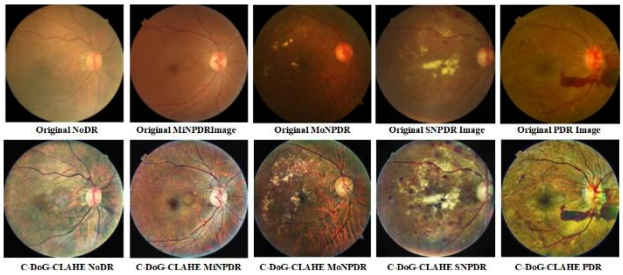

Figure 6: Per-class C-CLAHE-DoG enhanced samples

The characteristics of the training data significantly influence the retinal FI analysis in DL. An imbalanced distribution of data often results in the biasness toward the majority class, impacting final performance metrics. To address this issue, we employed geometric augmentation, in which geometric transformations, such as flipping and rotating at various angles, are applied, thereby increasing the number of samples in each class. For the combined (IDRiD and APTOS2019) dataset, each image is augmented using four geometric transformations which are rotation by 90° and 270°, horizontal and vertical flips. These transformed images are pooled with their original and enhanced counterparts to create the final enriched dataset. Representative samples are shown in Figure 7.

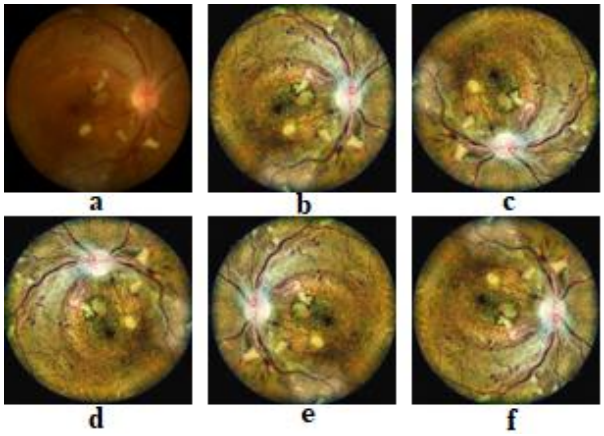

Figure 7: (a) Original image, (b) C-CLAHE-DoG image, (c) Rotate at 90°, (d) Rotate 270, (e) Horizontal Flip, (f) Vertical Flip.

For the DDR dataset, which exhibits pronounced class imbalance, augmenting all classes equally increased the imbalance issues. To mitigate this, we employed geometric minority oversampling, a class-aware augmentation technique that facilitates the balance of class distribution, enhancing the effectiveness of training DL models. This technique is used to selectively augment the underrepresented classes, ensuring equal participation in training. In DDR dataset, images from underrepresented classes, such as MiNPDR, SNPDR, and PDR, are augmented using the same four transformations and then merged with their enhanced counterparts. These transformations increase the dataset diversity and allow the model to generalize from a greater number of samples. Table 2 displays the combined and DDR dataset's pre- and post-augmentation samples.

Table 2: Pre- and post- augmented samples for combined (IDRiD+APTOS2019) and DDR datasets.

Classes	IDRiD+APTOS2019		DDR	
	Or.	**Aug.**	**Or.**	**Aug.**
Class1	395	2,370	630	3,150
Class2	1,167	7,002	4,477	4,477
Class3	286	1,716	236	1,180
Class4	357	2,142	913	4,565
Normal	1,974	11,844	6,266	6,266

To comply with the input requirements of the HARIC-Net architecture, bicubic interpolation is applied to resize all images to a consistent input dimension of 224 × 224 pixels. Bicubic interpolation is employed as it resizes more consistently than other techniques, such as nearest-neighbor interpolation, which can reduce image quality by adding blocky artifacts, and bilinear interpolation which performs a linear convolution and may cause blurring of high-frequency details. Bicubic interpolation uses a 4 × 4 neighborhood of surrounding pixels, resulting in smoother and more visually consistent outcomes. It effectively preserves fine vascular details, such as lesion boundaries, which are critical to retain after resizing, and enhances model's ability to identify important structures in FIs.

3.3 HARIC-Net for features extraction and classification

In this section, we provide a comprehensive description of our proposed HARIC-Net for DR screening, severity grading and comprehensive multi-class classification. The proposed architecture extracts heterogeneous low-level features like edges, contour boundaries, and vessel patterns using a Heterogeneous Features Aggregation (HFA) stem, as illustrated in Figure 8. In order to mitigate the feature loss that arises in simple feed forward stems, these features are fused at several levels within the stem. Furthermore, the model integrates multi-level feature extraction to ensure that the architecture extracts complex contextual information from large spatial regions while maintaining sensitivity to subtle local variations, and multiscale parallel filters to aggregate features across various resolutions, capturing comprehensive lesion details, especially enhancing high-frequency components that are rich in lesion representations.

Convolutional layers implement learned linear spatial filters followed by normalization and nonlinearity. For a single input channel, the forward convolution at spatial location (i, j) with an l × l kernel H is:

$$y_{i,j} = \sum_{u=1}^{l}\sum_{v=1}^{l} H_{u,v}x_{i+u-1,j+v-1} \tag{8}$$

and for a multi-channel filter $W \in \mathbb{R}^{\mathrm{l}\times\mathrm{l}\times \mathrm{C}_{in}\times \mathrm{C}_{out}}$ the output channel c' is :

$$Y_{i,j,c'} = \sum_{c=1}^{\mathrm{C}_{in}}\sum_{u=1}^{l}\sum_{v=1}^{l} W_{u,v,c,c'}H_{u,v}x_{i+u-1,j+v-1,c} + b_{c'} \tag{9}$$

Output spatial dimensions follow the usual relation $H_{out} = \left\lfloor\frac{H-l+2p}{8}\right\rfloor + 1$ for kernel size l, padding p, and stride s. HARIC-Net utilize small filters, specifically 1 × 1, 3 × 3 and 5 × 5 kernels to strikes a balance between classification performance and computational efficiency, and each convolutional layer is followed by batch normalization, which normalizes the activations to ensure stable output distributions. The Swish activation function is employed in HARIC-Net to mitigate information loss during the learning process, preserving the representational capacity of the network. In the proposed architecture, the resized input FI $x \in \mathbb{R}^{224\times224\times3}$ is passed through the HFA stem that extracts the rich and discriminative features using 3 × 3 and 5 × 5 convolutions. Unlike conventional feed forward stems that use shallow feed-forward layers, this HFA stem extracts and downsamples the rich semantic FMs using the hierarchal multi-branch pathways, and fuse them at varying depths within the stem to preserve both fine-grained and coarse contextual features. The HFA stem processes the colored FI of dimensions 224 × 224 × 3 (Width, Height, Color Channels) through a series of convolutional layers to extract essential spatial and feature representations from the input image. The first convolutional layer (Conv. L1) utilizes 64 filters of kernel size 3 × 3 with a Stride (S) of 1, to generate 224 × 224 × 64 FMs. These FMs are then passed through a max pooling layer having a 2 × 2 kernel and S=2, to reduce the spatial dimensions from (W × H) to (W / 2) × (H / 2). This results in FMs of size 112 × 112 × 64. These downsampled FMs are concatenated with the output from Conv. L2, which applies 32 filters of size 5 × 5 with $S = 2$ on the colored FI, yielding additional feature representations. Following this, the concatenated FMs are processed through Conv. L3 and Conv. L6. Conv. L3 applied 96 filters of size 3 × 3 with $S = 1$, followed by a max pooling layer to reduce the FM size to 56 × 56 × 96. These 56 × 56 × 96 FMs are fused with the output from Conv. L4, which applies 32 filters of size 5 × 5 with $S = 4$ on the input FI. The fused FMs are then passed through Conv. L5, which utilizes 128 filters of size 3 × 3 with S=1 followed by a max pooling layer, and generate 28 × 28 × 128 FMs. The 28 × 28 × 128 FMs are concatenated with the output from Conv. L6, which used 32 filters of size 5 × 5 with $S = 4$, resulting in the final output dimension of 28 × 28 × 160 FMs. Through heterogeneous aggregations, this multi-scale feature extraction captures low-level representations like edges and textures, allowing the network to learn robust and discriminative patterns from the FIs, that are essential for accurate DR diagnosis.

The output (28×28×128) from the HFA stem passes through the first SE attention block, which adaptively re-defines the channel responses to refine the FMs and highlight the features most influential for lesion

identification. Let $x \in \mathbb{R}^{H \times W \times C}$ be the input FMs, where H and W correspond to the spatial dimensions and and C is the number of channels. Throughout the network, 32 samples are processed together in a single pass. A channel descriptor is generated by the GAP operation, which collapses the spatial dimensions ($H \times W$) for every sample in the batch. While keeping the initial number of channels C, this squeeze operation aggregates spatial information to produce a 1 × 1 descriptor. Subsequently, the descriptor is processed through a Fully Connected (FC) layer and a ReLU nonlinearity, that utilizes only the positively impacting neurons, to reduce the representation's dimensionality to 1 × 1 × (C/r), where reduction rate is $r = 8$. As a resultant, an enhanced representation of the feature space is produced. This operation can be expressed mathematically as follows:

$$s = \sigma\big(W_2\delta(W_1 z)\big), \quad W_1 \in \mathbb{R}^{\frac{C}{r} \times C}, W_2 \in \mathbb{R}^{C \times \frac{C}{r}} \tag{10}$$

where W_1 and W_2 are the learnable weight matrices, $\sigma(\cdot)$ is the sigmoid, and $\delta(\cdot)$ is the ReLU activation. The output is then mapped back to the initial channel size. This is accomplished by passing it through a sigmoid activation function following a second FC layer. This results in channel-wise attention weights which are element-wise multiplied to the original FMs. This recalibrates the FMs which maintains the spatial resolution required for later local processing in the network and emphasize diagnostically essential channels.

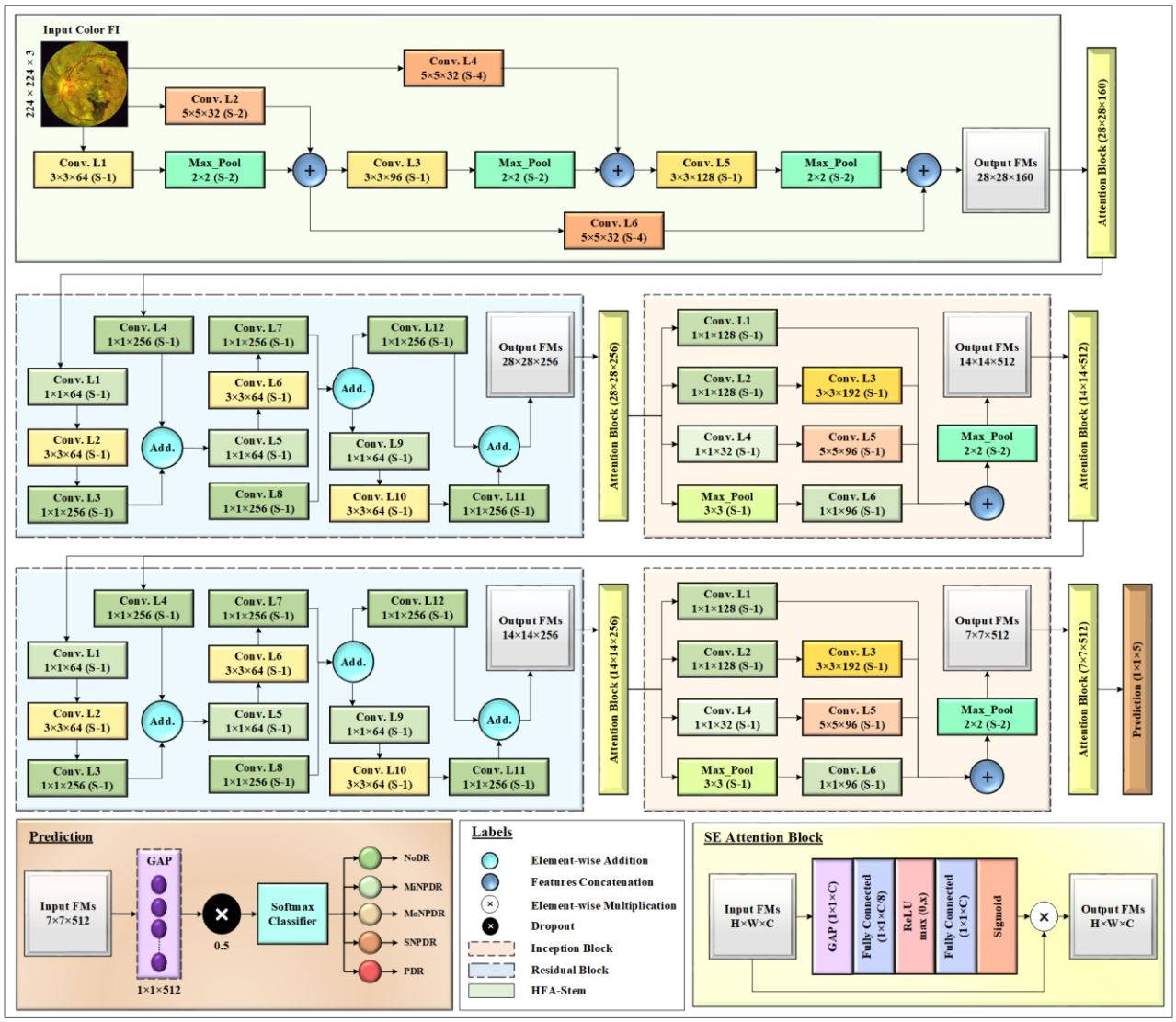

Figure 8: Proposed HARIC-Net's architecture diagram

SE Attention block refines the FMs without reducing the size of the images or the quantity of FMs. To capture sophisticated, multi-level features, the refined 28 × 28 × 128 FMs are subsequently passed through the first Residual Connection Block (RCB). The following is the formulation of the residual learning mechanism:

$$y = F\,(x; \{W_i\}) + W_s x, \tag{11}$$

where W_s is a linear projection (a 1 × 1 convolution) that can be applied when the input and output channel dimensions differ, and F is the residual function (i.e., the stacked convolutional layers). RCB is used as it can reduce the gradient vanishing problem and promote effective information flow in deep networks. Residual Block 1 begins by reducing the input FMs into a lower-dimensional space using 64 filters of size 1 × 1 (Conv. L1). A subsequent Conv. L2 of size 3 × 3 with 64 filters is applied, followed by Conv. L3 of size 1 × 1 with 256 filters, which restores the channel depth while maintaining the spatial dimensions. In parallel, Conv. L4 (1 × 1 × 256) is applied on the refined 28 × 28 × 128 FMs, and the output is added with the resultant FMs produced by Conv. L3. This sequence represents a residual unit, and is stacked three times. In particular, the added FMs pass through Conv. L5, Conv. L6, and Conv. L7, applying (1 × 1 × 64), (3 × 3 × 64) and (1 × 1 × 256) convolutions, respectively, and the output is then added to the FMs produced by a parallel Conv. L8 (1 × 1 × 256). Similarly, these added FMs containing more diverse features, pass through Conv. L9, Conv. L10, and Conv. L11, applying (1 × 1 × 64), (3 × 3 × 64) and (1 × 1 × 256) convolutions, respectively, and are added with the output FMs produced by Conv. L12 (1 × 1 × 256). After adding the FMs at the end of third residual unit, the final output of residual block 1 becomes 28 × 28 × 256 FMs. The stride value was kept at $S = 1$ throughout the RCB. This structure progressively captures deeper, more abstract representations.

In the RCB, each residual unit incorporates a shortcut connection that allows the input to bypass the intermediate layers. When the input and output dimensions differ, a 1×1 convolution is applied to the shortcut connection, aligning the dimensions before addition. This RCB architecture facilitates smoother backpropagation, enabling effective training of deeper networks. The semantically rich feature tensor of size 28 × 28 × 256 from Residual Block 1, is then passed through Attention Block 2 for further refinement. The attention block again emphasizes the valuable features and passes the FMs to Inception Block 1 to capture the multi-scale representations of DR. The proposed Inception-like block processes the incoming FMs through four parallel pathways. The outputs of these pathways are then concatenated along the channel dimension:

$$Y = [Y_1, Y_2, Y_3, Y_4], \tag{12}$$

where $Y \in \mathbb{R}^{\mathrm{H} \times \mathrm{W} \times (C_1, C_2, C_3, C_4)}$, with H and W representing the spatial dimensions, and C_1, C_2, C_3, C_4 denotes the number of output channels by each parallel pathway. Each branch of proposed Inception-like block is designed to extract information at various spatial granularities. Inception-like block process the input FMs of size 28 × 28 × 256, through four parallel convolutional branches. The first branch act as a dimensionality reducer. Conv. L1 in inception block extracts fine-details and local patterns by applying 1×1 convolution with 128 filters, while maintaining computational efficiency. Second branch applies Conv. L2 (1×1×128) followed by a 3×3 convolution (Conv. L3) with 192 filters to capture mid-level spatial representations and edge-like structures. These features are associated with moderate lesion regions, and help in capturing more complex patterns. The third branch applies a 1×1 bottleneck convolution (Conv. L4) using 32 filters, followed by a Conv. L5 of kernel size 5×5 with 96 filters. Given this configuration, the network responds to more global and wide-ranging retinal patterns, such as exudates or large hemorrhages, which are indicative of later stages of DR. In the fourth branch, dominant spatial features are preserved using a 3×3 max-pooling, and channel depth is restored and the pooled activations are integrated using a 1×1 convolution (Conv. L6) with 96 filters. Concatenating the outputs from all four branches results in multi-scale features aggregated FMs with a size of 28 × 28 × 512. Convolutions and max pooling throughout the parallel branches used $S = 1$. A subsequent 2×2 max-pooling layer with $S = 2$ downsamples the concatenated FMs to 14 × 14 × 512. This produces a richly descriptive tensor of decreased spatial dimensions having valuable information. Inception-like block captures fine, mid-level, and coarse features in parallel, to generate robust, discriminative representations across various DR stages, enabling the model to differentiate subtle abnormalities, such as microaneurysms, hemorrhages and exudates.

The output of Inception Block 1 (14 × 14 × 512) is passed through the attention block 3 to emphasize valuable features. These refined FMs are then fed into the Residual Block 2, where residual units progressively capture deeper and more semantically rich features, and produce a feature tensor of size 14 × 14 × 256. These semantically enriched FMs are passed through attention block 4 for further refinement. Attention block 4 keeps the enriched features and pass the FMs to Inception Block 2, which employs the same four parallel branch architecture as inception block 1, producing an output of 14 × 14 × 512 FMs after concatenation. A subsequent max-pooling layer in inception block 2 reduces the spatial dimensions of these FMs to 7 × 7 × 512, and pass them to Attention Block 5 for final channel gating. Attention block 5 refines and feed the FMs to the GAP layer, which is formulated as follows:

$$z_c = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} U_{i,j,c} \qquad c = 1, \ldots, 512. \tag{13}$$

The GAP head reform the spatial dimensions to a 1 × 1 × 512 feature vector z. Compared to an FC layer, a GAP layer efficiently reduces each channel and lowers the number of parameters, ensurring that a thorough multi-level and multi-scale feature representation facilitates the decision-making process. A dropout with a probability of $p = 0.5$ to the pooled vector is employed for regularization during training. The activations are scaled in accordance with the sampled binary mask $m \sim Bernoulli\,(1 - p)$ to approximate model averaging. This dropout ensures that, in the output layer, 50% of the input neurons are randomly set to 0 during each training

step. Dropout imposes a regularization penalty, to prevent the biasness towards specific features and improve generalization to unseen data. Finally, the dense layer computes the logits $Z \in \mathbb{R}^K$, one for each DR class. The Softmax classifier converts these logits into probabilities:

$$p_i = \frac{\exp(z_i)}{\sum_{j=1}^{k} \exp(z_j)} \qquad i = 1, \ldots\ldots, K. \tag{14}$$

HARIC-Net is a novel architecture, designed by utilizing canonical CNN building blocks, such as convolutions, pooling, bottleneck residual units, inception-style multi-branch fusion, GAP layer, dropout, and Softmax. This design is proposed to capture the diverse lesion cues of DR, and by combining attention-driven channel reweighting, HARIC-Net effectively focuses on diagnostically relevant channels and spatial patterns while maintaining computational efficiency.

3.4 The activation function

Activation function introduce nonlinearity in CNNs, that enables them to model complex mappings. Rectified Linear Unit (ReLU) is the most widely used activation function, that enabled the fully supervised training of deep neural networks. While ReLU is simple and effective, it causes "dead" neurons, when inputs fall in the non-active region. Mathematically, the ReLU function is defined as:

$$ReLU(x) = max(0, x) \tag{15}$$

Its elementwise derivative is:

$$\frac{d}{dx} ReLU(x) = \begin{cases} 1, & x > 0, \\ 0, & x < 0, \end{cases} \tag{16}$$

The zero derivative for negative inputs is the source of the "dead neuron" problem, meaning that once a unit outputs strictly negative pre-activation values, it may cease to update during gradient descent. The logistic sigmoid activation is another classical nonlinearity, defined as:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{17}$$

The sigmoid function maps any real input $x \in R$ to the interval $(0, 1)$. It's derivative $\sigma'(x) = \sigma(x)(1 - \sigma(x))$, is convenient for backpropagation. Despite being smooth and bounded between $(0, 1)$, sigmoid continues to suffer saturation for large $| x |$, which results in vanishing of gradient in deep networks.

To resolve these problems, the Swish activation function was introduced [43], which enabled the smoother flow and deeper feature representations. It is a non-monotonic function that allows small negative outputs and offers a gradual gradient change. Swish adapts itself based on incoming information, and reduce reliance on single-direction paths. Swish is created by gating the linear identity x with the sigmoid $\sigma(x)$ and is given by:

$$swish(x) = x \cdot \sigma(x) \tag{18}$$

where the sigmoid $\sigma(x)$ function is multiplied by x. The derivative of Swish can be written as:

$$\frac{d}{dx} swish(x) = \sigma(x) + x \cdot \sigma(x)(1 - \sigma(x)) \tag{19}$$

For a wide range of x, including negative inputs, this derivative stays nonzero. This lowers the possibility of inactive neurons and aids in preserving a smooth gradient flow during backpropagation.

The output FMs generated by the HFA stem using the ReLU and Swish activation functions are illustrated in Figure 9. Due to its non-smooth nature, the output of ReLU activation displays multiple blacked-out regions, which indicates the presence of dead neurons in the early layers. Swish activation, on the other hand, produces output FMs that are noticeably denser and smoother. By emphasizing on vessels and lesion-like patterns, Swish aids in displaying richer local structures with lesser non-zero neurons.

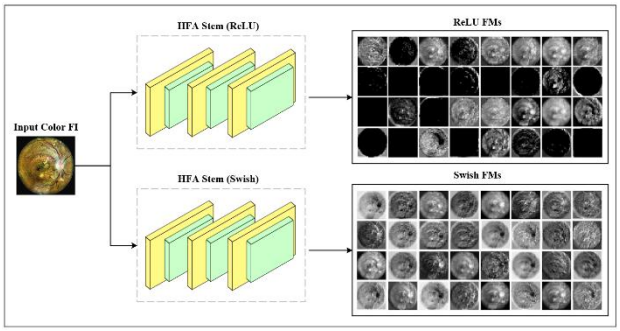

Figure 9: Impact of ReLU and Swish activations on the output of HFA-stem.

3.5 The loss function

In HARIC-Net, both binary and multi-class classification tasks are learned via supervised classification losses. While categorical cross-entropy is the conventional choice for mutually exclusive classes, it is known to be suboptimal in the presence of severe class imbalance. For a multi-class problem with C classes, the standard cross-entropy loss is defined as:

$$L(y, \hat{p}) = -\sum_{i=1}^{C} y_i \; log(\hat{p}) \tag{20}$$

where $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_C)$ is the one-hot encoded indicator for the true class. In the sparse implementation, the integer label is converted internally to its corresponding one-hot vector, and the same formula is applied. This approach treats each example equally, and as a result, the large majority classes dominate the gradient signal, leading to underfitting in minority classes. This underrepresented the

clinically significant severity levels. To mitigate this issue, we replaced the standard cross-entropy with a focal loss, which down-weights well-classified examples and concentrates learning on hard, misclassified samples [44]. Focal loss is specifically designed to assign greater significance to classes with fewer samples.

For a general C-class problem with N training samples, let $y_{ij} \in \{0,1\}$, indicates whether sample i belongs to class j, and let $\hat{p}_{i,j} = f_j(x_i;\theta)$ be the predicted probability for class j produced by the network $f_j(.;\theta)$. The multi-class focal loss used in our experiments is defined as:

$$L_{focal} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} \beta_j \left(1 - \hat{p}_{i,j}\right)^{\alpha} y_{ij} \log\left(\hat{p}_{i,j}\right) \quad (21)$$

where $\alpha \geq 0$ is the focusing parameter that controls how strongly the loss down-weights easy examples (larger α means more focus on hard examples), and $\beta_j > 0$ is a class-specific weight that compensates for unequal class frequencies (larger β_j increases the penalty for misclassifying class j). Intuitively, the $(1 - \hat{p}_{i,j})^{\alpha}$ term reduces the contribution of examples already predicted with high confidence (easy cases), while β_j rebalances per-class importance, ensuring that underrepresented severity levels exert greater influence during training.

For the binary case ($C = 2$) equation 21 simplifies to the familiar binary focal loss form, applied to the positive and negative class probabilities. When $\alpha = 0$ and $\beta_j = 1$ for all j, L_{focal} recovers the standard categorical cross-entropy. By using focal loss, HARIC-Net focuses more effectively on difficult examples and achieves more balanced grading performance across severity levels compared to training with standard cross-entropy, especially when the datasets exhibit pronounced class imbalance.

3.6 Model explainability via Grad-CAM

Among various XAI techniques, Grad-CAM is a widely used method that generates heatmaps to visually interpret CNNs by highlighting the image regions that contribute significantly to decision-making. Grad-CAM visually validates the focal points of the model's attention, confirming that the model is focusing on meaningful patterns. In HARIC-Net framework, Grad-CAM is employed to produce class-specific localization maps, to identify the regions within FIs that are most influential for a given prediction. Grad-CAM is computed on the final convolutional block immediately preceding the GAP layer in HARIC-Net architecture, generating visual explanations that indicate which retinal structures (e.g., microaneurysms, hemorrhages, exudates) strongly influence the model's output.

Formally, let $FM^k \in \mathbb{R}^{H \times W}$ denote the $k-th$ FM of the selected convolutional layer, and let y^c be the network's pre-softmax score for target class c. Grad-CAM computes channel-wise importance weights by taking the partial derivatives of y^c with respect to each spatial location of the FMs. Specifically, the weight α_k^c associated with channel k is obtained by spatially averaging the gradients over the FM domain:

$$\alpha_k^c = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\frac{\partial y^c}{\partial FM_{i,j}^k}. \quad (22)$$

In essence, α_k^c measures how sensitive the class score y^c is to changes in the activations of channel k. Large positive values indicate that increasing activations in that channel tends to increase the score for class c. The class-discriminative localization map (the Grad-CAM heatmap) is then formed by a weighted linear combination of the FMs, followed by a ReLU nonlinearity to retain only features that positively influence the target class:

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c FM^k\right). \quad (23)$$

Applying ReLU removes negative contributions (features that suppress the class score) and results in a coarse spatial map that highlights regions supporting the prediction for class c.

In practice, the resulting $H \times W$ heatmap is upsampled (e.g., by bilinear interpolation) to the original input image resolution and typically overlaid on the input FI to produce an interpretable visualization. These overlays allow qualitative assessment of whether the network attends to clinically meaningful retinal structures and can be used for case review by clinicians or to guide further model refinement.

3.7 Performance evaluation metrics

We used a number of performance measuring metrics, which were derived from the Confusion Matrix (CM), with an emphasis on precision, recall, and the F1-score, to assess the performance of the HARIC-Net in comparison to other baseline SOTA classifiers. Particularly when considering practical applications like DR diagnosis, these metrics aid in determining the efficacy of classifiers. Let $\mathrm{CM} \in \mathbb{N}^{n \times n}$ denote the CM for an n-class classification problem, where $\mathrm{CM}[i, j]$ represents the number of samples, having true label is class i and are predicted as class j. The standard CM-derived quantities for class i can be calculated as:

- True positives (TPs): $TP_i = \mathrm{CM}[i, i]$.
- False positives (FPs): $FP_i = \sum_{j=1, j\neq i}^{n} \mathrm{CM}[j, i]$.
- False negatives (FNs): $FN_i = \sum_{j=1, j\neq i}^{n} \mathrm{CM}[i, j]$.

Precision measures the TP predictions relative to all positive predictions made by model, whereas, recall evaluates the model's ability to correctly identify all TP samples. The F1-score, defined as the harmonic mean of precision and recall, is used evaluating model performance, in case of imbalanced datasets. Accuracy

measures how many accurate predictions a model makes in comparison to how many predictions it makes overall. It calculates the classification effectiveness of the entire model. These metrices are given as:

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \tag{24}$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \tag{25}$$

$$F1_i = \frac{2 \times \text{Precision}_i \times Recall_i}{\text{Precision}_i + Recall_i} \tag{26}$$

$$Accuracy = \frac{\sum_{i=1}^{n} CM\,[i,i]}{n} \tag{27}$$

These metrics allow for a more thorough evaluation of the proposed architecture's ability to manage unbalanced datasets while maintaining accuracy in necessary diagnostic tasks.

4 Experimental results and analysis

The NVIDIA Tesla T4 with 40 streaming multiprocessors, 6 MB shared L2 cache, and 16 GB GDDR6 memory, was utilized for training and evaluation. The GPU has a base clock of 1.59 GHz and can achieve a peak performance of approximately 8.1 TFLOPS in FP32 and 65 TFLOPS in FP16, and a maximum memory bandwidth of about 300 GB/s. Such a hardware setup meets the requirement of computational power and memory bandwidth to train HARIC-Net and other baseline models efficiently to compare these models with each other.

The batch size was maintained at 32 in all the experiments, to tradeoff between the GPU memory and stable gradient estimates. Each of the models used was trained over 60 epochs to provide sufficient time to learn feature representations. AdamW was used, which combines adaptive moment estimation and decoupled weight decay regularization. Its better convergence and ability to address overfitting, helps it to be chosen in the proposed architecture. Same protocols such as consistent data splits, preprocessing, and augmentation pipelines, and random-seed initializations, were used to train HARIC-Net and other baseline CNNs to reduce variability across architectures. This uniform training and selection process enabled equitable contrast of HARIC-Net with state of the art baselines.

4.1 Ablation study using IDRiD and APTOS2019 dataset

An ablation study is conducted to illustrate the effectiveness of the modules integrated in our proposed framework. We performed experiments on the IDRiD and APTOS2019 datasets, to gain insights about the impact of domain-shift challenges, augmentation and the SE-style attention blocks, on 2-class screening and 4-class severity grading of DR. The results provide understanding towards each component's contribution, to help deepen our understanding about the overall architectural behaviour. All samples underwent RoI segmentation and C-CLAHE-DoG enhancement pipeline prior to ablation evaluation. The enhancement pipeline combining CLAHE and DoG has been seen to improve lesion visibility in retinal FIs [22]. The test-set results are summarized in Table 3.

The proposed HARIC-Net achieved screening and grading accuracies of 83.33% and 46.47%, respectively, on the IDRiD dataset, reflecting the limited number of annotated samples in IDRiD for training. Insufficient data reduces the ability to learn reliable severity discriminators. In contrast, the APTOS2019 dataset, containing more training samples, produced 98.28% and 56.72%, screening and grading accuracies, respectively. The larger sample size, which offers better coverage of intra-class variability and enables more stable generalization, is primarily responsible for these improved results.

We combined the two datasets to expose the model to a diverse training and testing environment in order to analyse the domain shift issues. The screening and grading accuracy of the combined dataset were 98.08% and 55.75%, respectively. The combined dataset performs slightly lower than the APTOS2019 dataset by itself, because of the real-world inter- and intra-class variations introduced by the inclusion of IDRiD dataset. To solve this problem, we evaluated the impact of the augmentation towards generalizability. On the IDRiD dataset, screening and grading performances considerably improved up to 95.47% and 69.92%, respectively, proving that the geometric augmentations enhance the generalizability and decrease overfitting. Augmented APTOS2019 dataset achieved a screening accuracy of 98.36%, and a grading accuracy of 75.03%, which showed that such variability enables a fine-grained severity discrimination. Applying the geometric augmentation on the combined dataset also enhanced the grading and screening rates to 75.45% and 98.44%, respectively, highlighting that augmentation mitigates the challenges related to the heterogeneity of datasets and enhances the ability of the network to discriminate between meaningful changes.

Table 3: Ablation study results showing the impact of HARIC-Net's variants.

Model	IDRiD	APTOS2019	Augmentation	Attention Block	Screening Accuracy (%)	Grading Accuracy (%)
HARIC-Net V1.0	✓	✗	✗	✗	83.33	46.47
HARIC-Net V2.0	✗	✓	✗	✗	98.28	56.72

HARIC-Net V3.0	✓	✓	✗	✗	98.08	55.75
HARIC-Net V4.0	✓	✗	✓	✗	95.47	69.92
HARIC-Net V5.0	✗	✓	✓	✗	98.36	75.03
HARIC-Net V6.0	✓	✓	✓	✗	98.44	75.45
HARIC-Net Final	✓	✓	✓	✓	99.27	75.90
✗: Indicates exclusion of a component; ✓: Indicates inclusion of a component;						

In order to emphasize on the diagnostically critical features and focus on the important lesions, we used SE-attention block that enhanced screening and grading accuracy to 99.27% and 75.90%, respectively. This aligns with the behaviour of the attention block that inhibits the noisy and irrelevant areas and promotes the informative features. The joint effects of the addition of these elements provides significant increases in severity grading and coarse detection.

4.2 Performance evaluation of proposed HARIC-Net on combined dataset

The proposed framework generalized better as compared to the SOTA CNNs in the screening task, and accurately labelled 2,218 cases as TP out of 2,228 cases of DR, with only 10 cases being classified as FNs, as illustrated by the respective CM in Figure 10.

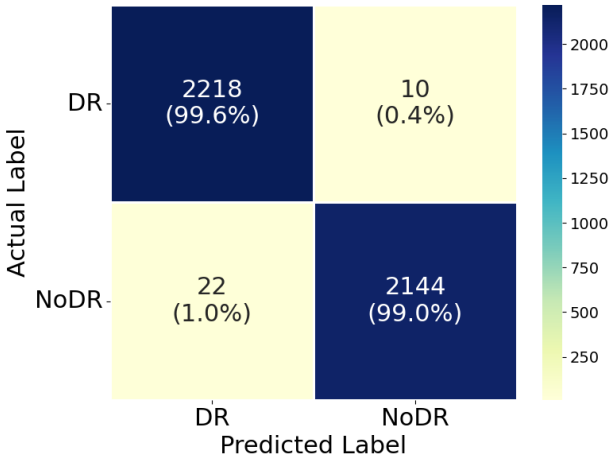

Figure 10: Test-set results of HARIC-Net for 2-class screening of DR.

The screening was carried out as it involves detection between healthy and diseased samples indicating the presence of disease. Out of 2,166 NoDR cases, HARIC-Net classifies only 22 cases as FP, and 2,144 cases as True Negatives (TNs), This gives a sensitivity (recall) of 99.26%, precision of 99.02%, F1-score of 99.28% and an overall accuracy of 99.27%, with low FN (~0.45%) and FP (~1.02%) values, proving that HARIC-Net reliably flags diseased samples. Both high sensitivity and high precision are useful especially in clinical screening since they decrease the risk of missed pathological cases.

Although screening can diagnose DR, it lacks in differentiating between the different stages of lesions. HARIC-Net is a solution to this limitation, where the severity of four different classes of tumors is graded. Problems with inter-and intra-class differences, and large imbalances in classes, compounded by the small number of samples, pose a challenge in differentiating lesion severities. Nonetheless, HARIC-Net identifies the variations of the severity of lesions, and categorize them into four clinically meaningful stages of DR, with the accuracy of 75.90% on 2,648 test samples as depicted in the CM in Figure 11.

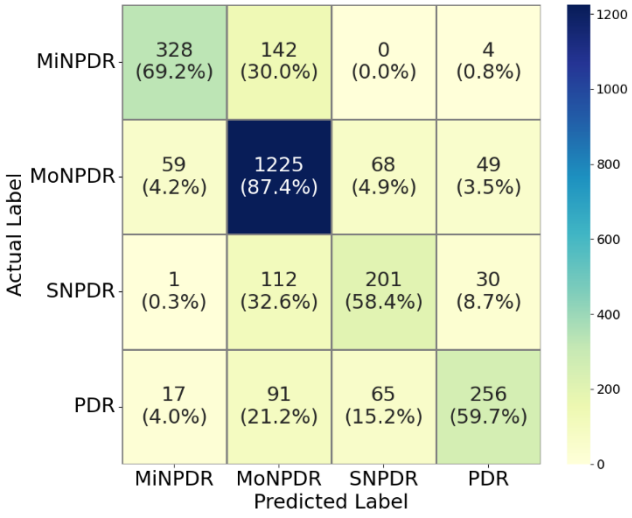

Figure 11: Test-set results of HARIC-Net for 4-class severity grading of DR.

MoNPDR presented highest class-wise recall of 87.44%, while MiNPDR, SNPDR and PDR achieved recall of 69.20%, 58.43% and 59.67%, respectively. Examining the off-diagonal entries reveals that due to their clinically overlapping characteristics, adjacent grades tend to get confused. Total 32.6% of SNPDR samples and nearly 30.0% of MiNPDR samples were miss-identified as MoNPDR. The clinical difficulty of early and progressive DR stages is further highlighted by the findings that 8.7% of the SNPDR samples were misclassified as PDR, while PDR cases were under-graded as MoNPDR (21.2%) or SNPDR (15.2%). These lesions share overlapping retinal features such as microaneurysms, dot-blot hemorrhages, and small exudates. Moreover, transitioning from SNPDR to PDR remains particularly ambiguous in single-field fundus imaging, even for expert graders.

Asymmetric performance across DR grades could result from class bias caused by the larger number of MoNPDR cases in the training data. A slight decrease in SNPDR and PDR recall indicates a risk of under-referral

for more advanced stages of the disease. The limited sample size and prominent class imbalance effects the model's ability to learn fine-grained distinctions between stages, leading to a decrease in overall statistical performance. However, by accurately classifying lesions across multiple severity levels despite the inherent challenges and overlapping features, the model provides valuable insights for clinicians, facilitating more precise diagnosis and treatment planning in DR. these findings suggest that the robust classification performance of HARIC-Net in screening and severity grading tasks is beneficial from the perspective of translational and clinical implementation.

4.3 Performance evaluation of proposed HARIC-Net on DDR dataset

The comprehensive multi-class grading experiments performed on the 5 classes of DDR dataset demonstrates that, while being computationally efficient, HARIC-Net achieves greater performances, and reveals class-specific strengths and weaknesses essential for clinical translation. A total of 3,931 images represented as NoDR, MiNPDR, MoNPDR, SNPDR, and PDR, were used for testing. Proposed HARIC-Net outperforms all the baseline SOTA models by achieving an overall accuracy of 80.76%. The inclusion of healthy class increases the sample size, and the model utilize the expanded diversity to improve the discrimination power between healthy and diseased samples, yielding a better overall classification performance. Per-class analysis reveals that the proposed architecture excels at identifying NoDR and PDR classes with recall of 91.55% and 89.82%, respectively, which suggests that only a few samples are missed, as illustrated in Figure 12. However, the performance for intermediate grades is a bit heterogeneous. MiNPDR achieves a recall of 74.60%, underrepresenting 14.0% of the mild cases as normal because of the asymptomatic nature of earl-stage lesions. Few samples are overcalled as moderate (5.6%) or proliferative (5.1%) disease.

Recall value of 67.34% for MoNPDR demonstrates a bit low yet balanced sensitivity, while underrepresenting 20.8% of the samples as normal, as early microaneurysms in peripheral retinal areas can be small and overlooked. Model accurately classifies 604 MoNPDR samples, and misclassifies few as milder (4.3%) or proliferative (5.9%) DR stage. SNPDR achieved 53.93% recall, demonstrating that most severe lesion samples are confused with PDR (25.4%) and MoNPDR (11.4%). SNPDR is characterized by retinal hemorrhages, and microvascular abnormalities, which may cause irregular retinal capillaries that can be mistaken as initial neovascularization stage of PDR.

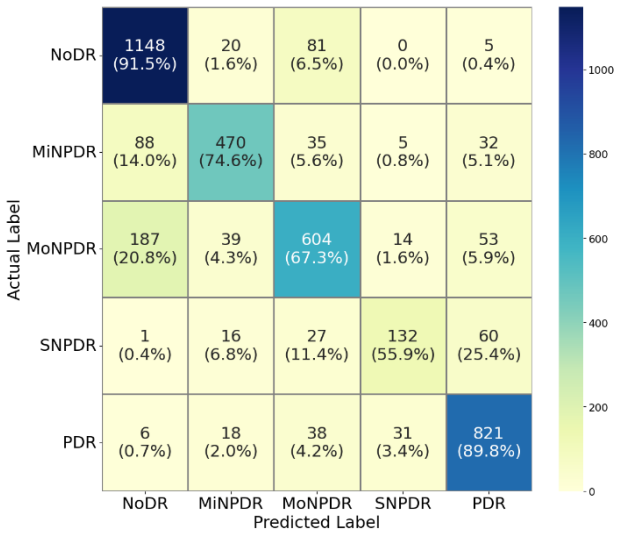

Figure 12: Test-set results of HARIC-net for 5-class classification of DR.

Investigating CM highlights that the confusion occurs between adjacent grades, which aligns with the clinical reality that progressive DR stages contain overlapping lesion structures, like microaneurysms, dot-blot hemorrhages, venous beading, and neovascularization, that are easy to overlook or confuse with other relative DR grades. The high recall of PDR and NoDR stages reduces the risk of misclassifying urgent cases and avoids unnecessary referrals in real-world clinical practices. However, the lower sensitivity for SNPDR and MoNPDR proposes a residual risk of under-referral for patients with vision-threatening stages of DR. Consequently, these results highlight that the proposed architecture demonstrates enhanced sensitivity for intermediate overlapping DR stages while having the low missing-rate for healthy and diseased cases.

4.4 Explainability insights on proposed framework's decisions

To emphasize on the interpretability and features contributing to the decision-making process of HARIC-Net, we computed Grad-CAM saliency maps from the network's final convolutional block. Heat maps were generated immediately preceding the GAP and classification layers, during testing on the DDR dataset, and overlaid on the corresponding test images. We choose Grad-CAM explainability over other methods because it effectively generates saliency heatmaps emphasizing the relevant regions of the image, and prominently localize disease-specific regions such as microaneurysms, hemorrhages, hard exudates, soft exudates, and cotton-wool spots, providing visual illustrations to ophthalmologists for better interpretation. Such visual explanations come in handy, as they can be matched with known diagnostic cues for strengthening the decision. Figure 13 presents representative test samples and their corresponding Grad-CAM heatmaps.

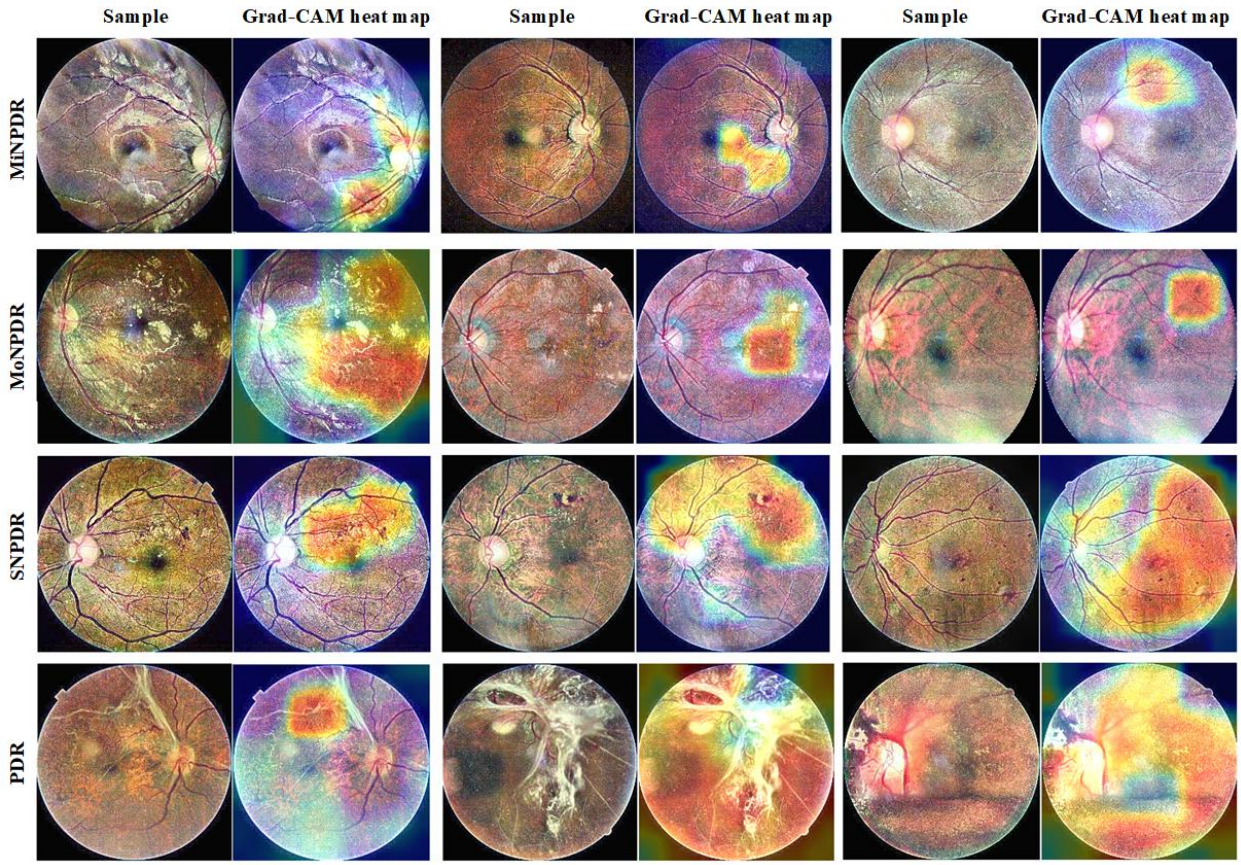

Figure 13: Grad-CAM visualizations for DR lesion classes

These visual representations can be used to investigate how the model pays attention to critical regions while training. Qualitatively, the saliency patterns are class-coherent and consistent throughout the sample set, emphasizing on the features utilized in decision-making.

The maps for MiNPDR are tightly localized, highlighting compact and high-intensity focal peaks, which correlates to the tiny punctate lesions adjacent to microaneurysms, that occurs in early stages. This indicates that the model focuses on subtle, vessel-like, high-frequency cues.

The attention regions for MoNPDR are larger and more diffuse, concentrating on the bright, clustered zones that corresponds to hard exudates and blot hemorrhages. This shift towards broader and contiguous regions, reflects the increased lesion count is influential for this stage.

In case of SNPDR, the heatmaps emphasize on pale patches and larger areas, which clinically correlates with cotton-wool spots and ischemic changes. The maps frequently trace disrupted vascular contours and areas of venous beading, reflecting the model's sensitivity to diffuse microvascular compromise.

The saliency is most extensive and intense for PDR, focusing on large hemorrhages fields and the regions having dense, high-frequency vascular structures, that are correspondent of neovascularization. The Grad-CAM saliency spans over multiple retinal quadrants, including the peripapillary zone, which aligns with clinical description of later proliferative stage.

The proposed novel C-CLAHE-DoG preprocessing pipeline highlights the diagnostically relevant features by emphasizing on local contrast and sharpening vessel edges, exudate boundaries, and neovascular activities, resulting in focused heatmaps and improved interpretability. Overall, the Grad-CAM results demonstrate that HARIC-Net utilizes the clinically meaningful retinal features while grading lesions, and provides an interpretable, evidence-based foundation for model refinement and clinical evaluation.

5 Discussion

Given that all SOTA CNNs and the proposed HARIC-Net, achieve near-ceiling performance in screening DR, the comparative analysis focuses on the 4-class severity grading and 5-class comprehensive classification tasks. Compared to a straightforward healthy/abnormal classification task, lesion classes present significant challenges due to issues like class imbalance, limited sample sizes, and overlapping patterns between stages. These factors often lead to suboptimal model performance, as models struggle to capture the subtle differences between classes. Moreover, in 5-class classification, with the inclusion of a healthy class alongside the four lesion classes, the number of samples

substantially increase which can cause the model to generalize more on the large sample set of normal class. This can lead to a suppression of the actual gains in performance that the model could achieve when differentiating the lesion severity levels. While the healthy class provides valuable data for generalization, the true challenge remains in fine-grained differentiation between the different stages of DR, which is often masked by the presence of a more dominant healthy class, which is why the 4-class severity grading offers a more rigorous and clinically applicable standard for assessing the usefulness of a model.

HARIC-Net provides a noticeably better equilibrium between precision and sensitivity in 4-class classification task on combined dataset, as shown by the comparative results in Figure 14.

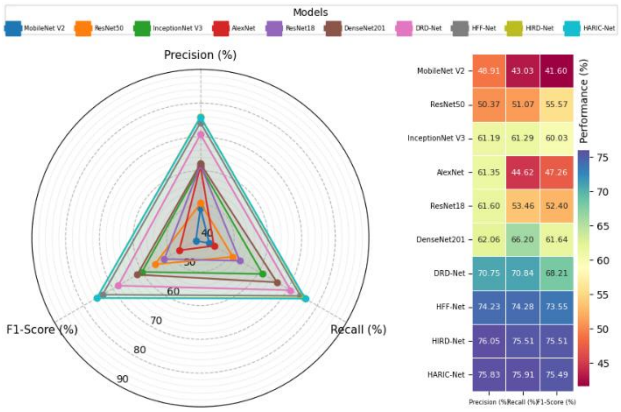

Figure 14: Comparative results for 4-class classification

Older and lightweight architectures perform lowest, i.e., MobileNet-V2 yields the lowest F1-score (41.60%) and recall (43.33%), demonstrating that how challenging it can be to capture fine-grained retinal lesions by using highly parameter-efficient architectures. ResNet50 and ResNet18, on the other hand, exhibit modest improvements with precision and recall falling between 50 to 62% and 51 to 53%, respectively, because of the residual skip connections. AlexNet produce a precision of 61.35% but a recall of 44.62%, showing a high degree of precision-recall imbalance. This is indicative of a conservative prediction bias which leads to fewer FPs prior to the missing lesions. However, those architectures as InceptionV3 and DenseNet201 that emphasize multi-scale representations and the reuse of dense features are more resistant to subtle lesions. InceptionV3 with a competitive recall of 61.29% performed moderately, while DenseNet201 achieved better recall of 66.20%.

DRD-Net shows a considerable improvement in its performance with the percentage of 70.75 and 70.84 of precision and recall, respectively, which proves the strong baseline of effective DR classification, whereas the relatively low F1-score proves there is a room for improvement. HFF-Net outperforms DRD-Net with balanced 74.23% precision, 74.28% recall, and 73.55% F1-score. Hierarchal approach improves generalization of the model results in a greater accuracy than the previous models. HIRD-Net has a precision of 76.05% and recall of 75.51%, which are great improvements compared to its predecessors. The precision-recall trade-off of the HIRD-Net reflects its effectiveness in minimizing both FPs and FNs, which is an important aspect of providing reliable DR diagnosis. The HARIC-Net has a precision of 75.83%, a recall of 75.91%, and an F1-score of 75.49%, thus presenting a significant improvement in relation to all baseline and hybrid CNNs. These findings highlight the effectiveness of the new architecture of HARIC-Net that integrates effective feature extraction methods and attention mechanisms. This makes the model an effective instrument, with a high recall rate, in detecting subtle lesions in the retina without overlooking important details, making it a very useful tool in 4-class severity grading.

In contrast, HARIC-Net generalizes to heterogeneous, real-world retinal images more robustly, when classified on DDR benchmark's healthy and lesion classes, as depicted in Figure 15. The extreme parameter efficiency and depthwise separable operations of MobileNet-V2 fails to capture the fine-grained, high-variation lesion patterns found across a diverse clinical cohort, as demonstrated by the model's precision (60.59%), recall (55.19%), and F1-score (56.07%).

AlexNet has a higher accuracy than MobileNet-V2, with a precision of 65.09%, recall of 63.04%, and F1-score of 63.74%. The sequential nature of feature extraction used in AlexNet with its more convoluted architecture, nevertheless, cannot easily detect fine-grained lesions. ResNet50 also takes advantage of residual connectivity to enhance feature propagation with the precision of 66.86%, recall of 62.12% and F1-score of 63.17%. With a lightweight residual network, ResNet18 has a higher precision of 67.75%, a recall of 64.71% and a higher F1-score of 65.71%, which is an incremental improvement over ResNet50. Multi-branch and multi-scale convolutions in InceptionV3 enable the network to achieve a much higher performance of 74.91% accuracy, 70.96% recall and 72.46% F1-score. Its capability in capturing spatially heterogeneous lesions signatures due to its groundbreaking network architecture achieves more robust lesion detection. DenseNet201 having dense connections and a feature reuse mechanism yields a precision of 78.63%, a recall of 78.70% and a F1-score of 78.27%.

DRD-Net has precision, recall, and F1-score of 79.18%, 78.40%, and 77.57%, respectively, which are slightly better than DenseNet201. The sophisticated feature extraction schemes and improved training mechanisms that are incorporated in the model lead to its high performance. HFF-Net outperforms DRD-Net with a precision of 79.44% and a recall and F1-score of 79.20% and 79.39%, showcasing that it can more easily handle the complexity of retinal images with its hierarchical feature fusion strategies. HIRD-Net gains 79.88% precision, 79.85% recall and 79.74% F1-score, with a higher sensitivity to small lesions and no compromise in the specificity. Lastly, HARIC-Net has a 79.55% precision, 75.84% recall, 77.28% F1-score and a total accuracy of 80.76%, which is very impressive in terms of overall performance in all metrics. Although the recall is low, the

overall high performance with a tradeoff of ~2 M parameters suggests that HARIC-Net enhances sensitivity to subtle lesions without compromising specificity, while being computationally efficient.

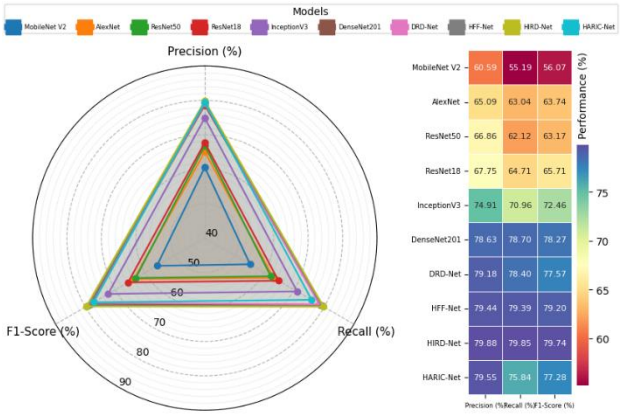

Figure 15: Comparative results for 5-class classification.

A clear trade-off between the number of parameters and practical effectiveness of proposed framework, SOTA baselines and other hybrid CNNs is revealed by the comparative analysis of classification accuracies and complexities, shown in Table 4.

MobileNet-V2 with a highly compact architecture (~3.4M parameters) only attains 56.3% 4-class and 64.0% 5-class accuracy, which suggests that vigorous pruning of the network does not preserve the high-frequency, subtle retinal lesion cues. At the other end, AlexNet only achieve 61.63% and 70.03% accuracies despite being having ~61M parameters, suggesting that merely adding more parameters is not beneficial. The ResNet50 (approximately 25.6 M parameters) has an accuracy of 59.78% on the 4-class and 70.54% on the 5-class classification task. Although it has high number of parameters, relatively low performance implies that its dense-residual connections do not provide the multi-scale feature processing needed to encode the full complexity of retinal lesions. The 4-class classification task and the 5-class classification task are slightly higher with ResNet18 of ~11.7 million parameters with an accuracy of 61.85% and 71.74%, respectively. InceptionV3 (~23.8 M parameters), which operates on features of varying scales through multi-branch kernels, has been shown to improve with 61.29% accuracy on the 4-class classification task and 76.88% on the 5-class classification task. DenseNet201, which has approximately 20.2 M parameters, attains accuracy rates of 66.20% and 78.70% on the 4-class and 5-class classification tasks respectively, is more sensitive to micro-lesions and can be more generally applicable to a variety of DR stages.

The recent lightweight hybrid designs show that specialized architectural features can be highly accurate using much fewer parameters. DRD-Net, with a size of approximately 3.7M parameters, has a high performance of 70.84% on a 4-class and 78.40% on a 5-class classification task, which demonstrates that smaller size could be beneficial with focused design. Similarly, HFF-Net using just about ~1.18M parameters adds both hierarchical features fusion and multi-level dense connections, resulting in 74.28% 4-class and 79.39% 5-class grading accuracy. This demonstration underscores that well-organized dense connections and multi-scale interconnection can see a very lightweight network perform better than bigger backbones. HIRD-Net of about ~4.8M parameters further illustrates the effect of the ability to capture morphological information. It includes attention and strategic features enhancement to focus on structural characteristics such as microaneurysms and hemorrhages with a 75.71% 4-class and 79.85% 5-class classification accuracy. Specifically, fine morphological cues can be extracted with selective filters and attention mechanisms to increase performance without enormous parameterization.

HARIC-Net achieves the best balance of accuracy and compactness. It achieves 75.90% 4-class severity grading and 80.76% 5-class comprehensive classification accuracy, the best of all the models, despite having less parameters of the order of 2M. HARIC-Net is more effective than computationally-demanding SOTA CNNs that prove their design choices. Its multi-level residual and inception-style blocks with targeted preprocessing exploit multi-scale features effectively. The high accuracy to parameter ratio implies that HARIC-Net occupies less memory space and can be inferred more quickly.

Table 4: Comparison of overall accuracies and model complexities, for 4-class and 5-class classification tasks, across various CNNs.

CNN Architectures	4-Class Classification Acc. (%)	5-Class Classification Acc. (%)	Parameters (Million)
MobileNetV2	56.30	64.00	~3.4
AlexNet	61.63	70.03	~61
ResNet50	59.78	70.54	~25.6
ResNet18	61.85	71.74	~11.7
InceptionV3	61.29	76.88	~23.8
DenseNet201	66.20	78.70	~20.2
DRD-Net [25]	70.84	78.40	~3.7
HFF-Net [27]	74.28	79.39	~1.18
HIRD-Net [22]	75.71	79.85	~4.8
HARIC-Net	75.90	80.76	~2

6 Conclusions and future work

Diabetes causes individuals to lose their vision to DR, and early identification is important in the prevention or delay of the onset of visual impairment. To minimize human error in the diagnosing process and allow real-time applications, an automated DR detection method is necessary. This research introduces a DL architecture named HARIC-Net, which appears to have a very small computational footprint and high-quality DR diagnosis with enhanced transparency. A retinal RoI cropping, channel-wise CLAHE enhancement (C-CLAHE), and DoG fusion combines to create a robust preprocessing pipeline, which improves the local contrast, and highlight fine-gained features. To retain the micro-level lesion cues and enhance broad contextual information, the architecture of the model integrates residual bottlenecks, inception-style multi-branch filters, and SE attention blocks. The network utilizes smooth Swish activation function, and a focal loss formulation to optimise the workflow of major blocks, which enables the model to learn the hard-to-classify, under-represented severity levels down-weighted by the well-classified examples. Moreover, the interpretation of the classification outcome is made easier by the visualization of Grad-CAM saliency maps, which provide a clearer spatial correspondence between the lesion structures. The saliency fields of the final convolutional block that are highly co-localized with the lesions in the retinal images exhibit a higher spatial precision when Grad-CAM maps are used.

To validate the effectiveness of our methodology, we conducted extensive ablation and comparative experiments on three publicly available benchmark datasets. The results of the experiment on IDRiD, APTOS2019 and DDR datasets, for screening, severity grading and comprehensive classification, are the indicative of better generalization ability of HARIC-Net, while differentiating among the various stages of DR. With just ~2 M trainable parameters, the model efficiently demonstrated competitive grading performances and clinical-grade screening performances in our experiments. HARIC-Net is particularly compatible with the use of hardware having limited resources on clinical translation due to its less footprint and high recall on proliferative cases. Moreover, it is best suited to be used in triage pipelines, in which a high level of sensitivity to those diseases that possess a risk to vision will be prioritized, whereas unnecessary referrals will be minimized.

For future development, we advocate following side-by-side development directions to attain clinical readiness: first, to enhance localization and class separability through the incorporation of lesion-level or quadrant-wise annotations and multi-field inputs to minimize ambiguity in individual images; second, to quantify and report runtime and resource footprints, carry out calibration studies, and enforce prospective external validation in order to assess real-world robustness and decision limits; third, to explore complementary modalities and modeling paradigms, such as multimodal fusion with clinical metadata and Optical Coherence Tomography; and forth, to include methods of quantitative explainability, including perturbation-based metrics, to evaluate the faithfulness of model explanations. This would entail assessing the effect of model confidence on varying key image regions as alterations in the image are gradually introduced by either deleting or adding the most significant pixels [44]. These metrics will give us information on the model relying on particular image characteristics and we will be able to understand the way the model makes decisions. Overall, this experimental study and results indicate that HARIC-Net is a highly effective, interpretable, and useful framework to grade severity level, particularly normal and proliferative, with high accurateness, indicating its potential as a scalable solution to improve the diagnosis of DR.

Acknowledgement

The authors conducted this research independently and were solely responsible for all aspects of the study, including data preprocessing, model development, experimentation, and analysis. No external funding or institutional support was received.

References

[1] N. Salamat, M. M. S. Missen, and A. Rashid, “Diabetic retinopathy techniques in retinal images: A review,” *Artif. Intell. Med.*, vol. 97, pp. 168–188, 2019, doi: https://doi.org/10.1016/j.artmed.2018.10.009.

[2] M. Nahiduzzaman *et al.*, “Diabetic retinopathy identification using parallel convolutional neural network-based feature extractor and ELM classifier,” *Expert Syst. Appl.*, vol. 217, May 2023, doi: 10.1016/j.eswa.2023.119557.

[3] World Health Organization (WHO), “Diabetes.”

[4] P. Lanzetta *et al.*, “Fundamental principles of an effective diabetic retinopathy screening program,” *Acta Diabetol.*, vol. 57, no. 7, pp. 785–798, Dec. 2020, doi: 10.1007/S00592-020-01506-8/TABLES/2.

[5] M. Mateen, T. S. Malik, S. Hayat, M. Hameed, S. Sun, and J. Wen, “Deep Learning Approach for Automatic Microaneurysms Detection,” *Sensors*, vol. 22, no. 2, Jan. 2022, doi: 10.3390/s22020542.

[6] S. Wang *et al.*, “Diabetic Retinopathy Diagnosis Using Multichannel Generative Adversarial Network with Semisupervision,” *IEEE Transactions on Automation Science and Engineering*, vol. 18, no. 2, pp. 574–585, Dec. 2021, doi: 10.1109/TASE.2020.2981637.

[7] D. A. da Rocha, F. M. F. Ferreira, and Z. M. A. Peixoto, “Diabetic retinopathy classification using VGG16 neural network,” *Research on Biomedical Engineering*, vol. 38, no. 2, pp. 761–772, Dec. 2022, doi: 10.1007/S42600-022-00200-8/TABLES/6.

[8] B. Menaouer, Z. Dermane, N. E. H. Kebir, and N. Matta, “Diabetic Retinopathy Classification Using Hybrid Deep Learning Approach,” *SN*

Comput. Sci., vol. 3, no. 5, Dec. 2022, doi: 10.1007/S42979-022-01240-8.

[9] J. R, V. Sivasubramanian, J. Prakash, and V. M, "Detection of Diabetic Retinopathy Using Convolutional Neural Networks," *ECS Trans.*, vol. 107, no. 1, pp. 13321–13328, Dec. 2022, doi: 10.1149/10701.13321ECST/XML.

[10] V. Vives-Boix and D. Ruiz-Fernández, "Diabetic retinopathy detection through convolutional neural networks with synaptic metaplasticity," *Comput. Methods Programs Biomed.*, vol. 206, p. 106094, Dec. 2021, doi: 10.1016/J.CMPB.2021.106094.

[11] M. H. Ashraf, F. Jabeen, H. Alghamdi, M. S. Zia, and M. S. Almutairi, "HVD-Net: A Hybrid Vehicle Detection Network for Vision-Based Vehicle Tracking and Speed Estimation," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 8, p. 101657, Dec. 2023, doi: 10.1016/J.JKSUCI.2023.101657.

[12] M. H. Ashraf, M. Ahmed, M. N. Mehmood, M. E. Qureshi, and H. Alghamdi, "MSHFF-Net: An Explainable Multi-Scale Hierarchical Feature Fusion CNN for Breast Cancer Diagnosis from Histopathological Images," in *2025 27th International Multitopic Conference (INMIC)*, Islamabad: IEEE, Jan. 2026, pp. 1–6.

[13] M. E. Qureshi, M. H. Ashraf, M. W. Arshad, A. Khan, H. Ali, and Z. U. Abdeen, "Hierarchical Feature Fusion With Inception V3 for Multiclass Plant Disease Classification," *Informatica*, vol. 49, no. 27, Jul. 2025, doi: 10.31449/inf.v49i27.8208.

[14] N. Gharaibeh, O. M. Al-Hazaimeh, A. Abu-Ein, and K. M. O. Nahar, "A Hybrid SVM NAÏVE-BAYES Classifier for Bright Lesions Recognition in Eye Fundus Images," *International Journal on Electrical Engineering and Informatics*, vol. 13, no. 3, 2021, doi: 10.15676/ijeei.2021.13.3.2.

[15] M. K. Yaqoob, S. F. Ali, M. Bilal, M. S. Hanif, and U. M. Al-Saggaf, "ResNet Based Deep Features and Random Forest Classifier for Diabetic Retinopathy Detection," *Sensors 2021, Vol. 21, Page 3883*, vol. 21, no. 11, p. 3883, Dec. 2021, doi: 10.3390/S21113883.

[16] N. Asiri, M. Hussain, F. Al Adel, and N. Alzaidi, "Deep learning-based computer-aided diagnosis systems for diabetic retinopathy: A survey," *Artif. Intell. Med.*, vol. 99, p. 101701, Dec. 2019, doi: 10.1016/J.ARTMED.2019.07.009.

[17] F. Li *et al.*, "Deep learning-based automated detection for diabetic retinopathy and diabetic macular oedema in retinal fundus photographs," *Eye 2021 36:7*, vol. 36, no. 7, pp. 1433–1441, Dec. 2021, doi: 10.1038/s41433-021-01552-8.

[18] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, Dec. 2014, [Online]. Available: https://arxiv.org/abs/1409.1556v6

[19] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," 2016.

[20] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," 2017. [Online]. Available: https://github.com/liuzhuang13/DenseNet.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2016. [Online]. Available: http://image-net.org/challenges/LSVRC/2015/

[22] M. H. Ashraf *et al.*, "HIRD-Net: An Explainable CNN-Based Framework with Attention Mechanism for Diabetic Retinopathy Diagnosis Using CLAHE-D-DoG Enhanced Fundus Images," 2025, doi: 10.3390/xxxxx.

[23] M. N. Hasan, M. E. R. Pial, S. Das, N. Siddique, and H. Wang, "DIA-VXNET: A framework for automated diabetic eye disease detection using transfer learning with feature fusion network," *Biomed. Signal Process. Control*, vol. 100, Aug. 2025, doi: 10.1016/j.bspc.2024.106907.

[24] R. Vij and S. Arora, "Modified deep inductive transfer learning diagnostic systems for diabetic retinopathy severity levels classification," *Biomed. Signal Process. Control*, vol. 99, Aug. 2025, doi: 10.1016/j.bspc.2024.106885.

[25] M. H. Ashraf, M. E. Qureshi, A. Khan, and M. Ahmed, "DRD-Net: Diabetic Retinopathy Diagnosis Using A Hybrid Convolutional Neural Network," *International Journal on Robotics, Automation and Sciences*, vol. 7, no. 2, p. 96, 2025, doi: 10.33093/ijoras.2025.7.2.9.

[26] A. Bin Ahmed and M. N. Mehmood, "Forecasting Solar Project Capacity Trends Using Predictive AI: A Policy-Oriented Analysis of Urban vs. Rural Solar Deployment," in *Proceedings of the Sixth International Conference on Digital Age & Technological Advances for Sustainable Development*, New York, NY, USA: ACM, May 2025, pp. 66–73. doi: 10.1145/3747897.3747909.

[27] M. H. Ashraf and H. Alghamdi, "HFF-Net: A hybrid convolutional neural network for diabetic retinopathy screening and grading," *Biomedical Technology*, vol. 8, pp. 50–64, Dec. 2024, doi: 10.1016/J.BMT.2024.09.004.

[28] F. M. J. M. Shamrat *et al.*, "An advanced deep neural network for fundus image analysis and enhancing diabetic retinopathy detection," *Healthcare Analytics*, vol. 5, Sep. 2024, doi: 10.1016/j.health.2024.100303.

[29] M. Herrero-Tudela, R. Romero-Oraá, R. Hornero, G. C. G. Tobal, M. I. López, and M. García, "An explainable deep-learning model reveals clinical clues in diabetic retinopathy through SHAP," *Biomed. Signal Process. Control*, vol. 102, Sep. 2025, doi: 10.1016/j.bspc.2024.107328.

[30] P. Venkatasubbu, A. V. Shreyas Madhav, K. Prakash, R. Rajaraman, and L. Bhaskara Rao, "Interpretable deep automation system for graded diagnosis of diabetic retinopathy," *Biomed. Signal Process. Control*, vol. 112, Feb. 2026, doi: 10.1016/j.bspc.2025.108490.

[31] S. J. SIDIQ and T. BENIL, "A lightweight transfer learning-based ensemble approach for diabetic retinopathy detection," *International Journal of Information Management Data Insights*, vol. 5, no. 2, Dec. 2025, doi: 10.1016/j.jjimei.2025.100372.

[32] N. M. Suganthi and M. Arun, "Diabetic retinopathy grading using curvelet CNN with optimized SSO activations and wavelet-based image enhancement," *Ain Shams Engineering Journal*, vol. 16, no. 1, Aug. 2025, doi: 10.1016/j.asej.2024.103239.

[33] K. Ashwini and R. Dash, "Improving Diabetic Retinopathy grading using Feature Fusion for limited data samples," *Computers and Electrical Engineering*, vol. 120, Dec. 2024, doi: 10.1016/j.compeleceng.2024.109782.

[34] D. B. Soomro *et al.*, "Automated dual CNN-based feature extraction with SMOTE for imbalanced diabetic retinopathy classification," *Image Vis. Comput.*, vol. 159, Aug. 2025, doi: 10.1016/j.imavis.2025.105537.

[35] S. Madarapu, S. Ari, and K. Mahapatra, "A multi-resolution convolutional attention network for efficient diabetic retinopathy classification," *Computers and Electrical Engineering*, vol. 117, Jul. 2024, doi: 10.1016/j.compeleceng.2024.109243.

[36] A. K. Singh, S. Madarapu, and S. Ari, "Diabetic retinopathy grading based on multi-scale residual network and cross-attention module," *Digital Signal Processing: A Review Journal*, vol. 157, Feb. 2025, doi: 10.1016/j.dsp.2024.104888.

[37] S. E. Abraham and B. C. Kovoor, "Dual-stage dynamic hierarchical attention framework for saliency-aware explainable diabetic retinopathy grading," *Eng. Appl. Artif. Intell.*, vol. 148, May 2025, doi: 10.1016/j.engappai.2025.110364.

[38] R. D S and K. S. Saji, "Hybrid deep learning framework for diabetic retinopathy classification with optimized attention AlexNet," *Comput. Biol. Med.*, vol. 190, May 2025, doi: 10.1016/j.compbiomed.2025.110054.

[39] P. Porwal *et al.*, "Indian Diabetic Retinopathy Image Dataset (IDRiD): A Database for Diabetic Retinopathy Screening Research," 2018, doi: 10.21227/H25W98.

[40] S. D. K. Maggie, "APTOS 2019 Blindness Detection," 2019, *Kaggle*. [Online]. Available: https://kaggle.com/competitions/aptos2019-blindness-detection

[41] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, and H. Kang, "Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening," *Inf. Sci. (N. Y).*, vol. 501, pp. 511–522, Oct. 2019, doi: 10.1016/j.ins.2019.06.011.

[42] T. A. Soomro *et al.*, "Impact of novel image preprocessing techniques on retinal vessel segmentation," *Electronics (Switzerland)*, vol. 10, no. 18, Sep. 2021, doi: 10.3390/electronics10182297.

[43] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for Activation Functions," Oct. 2017, [Online]. Available: http://arxiv.org/abs/1710.05941

[44] M. H. Ashraf, F. Jabeen, M. Waqar, and A. Kim, "HFA-Net: Explainable Multi-Scale Deep Learning Framework for Illumination-Invariant Plant Disease Diagnosis in Precision Agriculture," *Sensors*, vol. 26, no. 7, Apr. 2026, doi: 10.3390/s26072067.