Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.061150

ARTICLE

MVLA-Net: A Multi-View Lesion Attention Network for Advanced Diagnosis and Grading of Diabetic Retinopathy

Tariq Mahmood[1,2], **Tanzila Saba**[1], **Faten S. Alamri**[3,*], **Alishba Tahir**[4] and **Noor Ayesha**[5]

[1]Artificial Intelligence and Data Analytics (AIDA) Lab, College of Computer & Information Sciences, Prince Sultan University, Riyadh, 11586, Saudi Arabia

[2]Faculty of Information Sciences, University of Education, Vehari Campus, Vehari, 61100, Pakistan

[3]Department of Mathematical Sciences, College of Science, Princess Nourah bint Abdulrahman University, Riyadh, 84428, Saudi Arabia

[4]Shifa College of Medicine, Shifa Tameer-E-Millat University, Islamabad, 44000, Pakistan

[5]Center of Excellence in Cyber Security (CYBEX), Prince Sultan University, Riyadh, 11586, Saudi Arabia

*Corresponding Author: Faten S. Alamri. Email: fsalamripnu@gmail.com or fsalamri@pnu.edu.sa

Received: 18 November 2024; Accepted: 18 February 2025; Published: 26 March 2025

ABSTRACT: Innovation in learning algorithms has made retinal vessel segmentation and automatic grading techniques crucial for clinical diagnosis and prevention of diabetic retinopathy. The traditional methods struggle with accuracy and reliability due to multi-scale variations in retinal blood vessels and the complex pathological relationship in fundus images associated with diabetic retinopathy. While the single-modal diabetic retinopathy grading network addresses class imbalance challenges and lesion representation in fundus image data, dual-modal diabetic retinopathy grading methods offer superior performance. However, the scarcity of dual-modal data and the lack of effective feature fusion methods limit their potential due to multi-scale variations. This paper addresses these issues by focusing on multi-scale retinal vessel segmentation, dual feature fusion, data augmentation, and attention-based grading. The proposed model aims to improve comprehensive segmentation for retinal images with varying vessel thicknesses. It employs a dual-branch parallel architecture that integrates a transformer encoder with a convolutional neural network encoder to extract local and global information for synergistic saliency learning. Besides that, the model uses residual structures and attention modules to extract critical lesions, enhancing the accuracy and reliability of diabetic retinopathy grading. To evaluate the efficacy of the proposed approach, this study compared it with other pre-trained publicly open models, ResNet152V2, ConvNext, Efficient Net, DenseNet, and Swin Transform, with the same developmental parameters. All models achieved approximately 85% accuracy with the same image preparation method. However, the proposed approach outperforms and optimizes existing models by achieving an accuracy of 99.17%, 99.04%, and 99.24%, on Kaggle APTOS19, IDRiD, and EyePACS datasets, respectively. These results support the model's utility in helping ophthalmologists diagnose diabetic retinopathy more rapidly and accurately.

KEYWORDS: Diabetic retinopathy grading; retinal vessel segmentation; dual-modal; deep learning; attention mechanism; health risks

1 Introduction

The demand for eye health is increasing due to economic and social development and an aging population. However, high-quality ophthalmic resources are limited and unevenly distributed. Diagnosing diabetic retinopathy (DR) requires manual observation, which is time-consuming and heavily dependent on specialist expertise [1]. Integrating technologies like big data, artificial intelligence, and 5G can enhance early screening capabilities for eye diseases [2]. Research into efficient automatic diagnosis algorithms for DR provides the theoretical foundation for integrating AI with clinical applications, improving diagnostic accuracy and efficiency, promoting large-scale screening, and facilitating early DR prevention and treatment. Continuous improvement of segmentation and grading algorithms is necessary due to the complex pathological changes and similarities across severity levels of DR. Clinically, DR detection and severity grading are mainly performed by visual inspection using color fundus imaging. DR is classified internationally into five levels, ranging from no DR to the most severe proliferative DR [3,4]. Accurate diagnosis is challenging due to similar symptoms across DR stages.

In recent years, AI-driven methods have gained significant attention for medical image analysis and disease diagnosis, particularly in the segmentation of retinal pathological changes and automatic disease classification algorithms [5]. Accurate vessel segmentation is crucial for individualized DR treatment planning, providing clinicians with deeper insights and more precise diagnostic opportunities [6,7]. Vessel segmentation has the potential to broaden and improve DR research, providing scientific and effective treatment methods with significant implications for ophthalmology. However, balancing global and local information remains a challenge. This study designs a multi-scale retinal vessel segmentation method that fuses global and local information to effectively model complex vascular structures and improve segmentation accuracy for vessels of varying sizes [8].

Smart healthcare uses automatic classification algorithms to rapidly and accurately classify retinal diseases. The automatic approach extends medical resources and promotes effective vision screening, fundus screening technology, and early detection, diagnosis, and treatment at the grassroots level. It reduces disease burden and blindness rates, enhances prevention, diagnosis, and follow-up services, and supports large-scale eye disease screening. Automated grading aids doctors with diagnostic insights and treatment suggestions, improving diagnostic speed and accuracy [9]. However, the imbalance in fundus dataset distribution and the complexity of lesions in images limit the accuracy of automated DR grading. Multi-scale, attention-based, and multi-stage training strategies have been widely adopted to address these challenges. Vision Transformer-based models capture long-range dependencies between lesions, helping the encoder better understand medical image information [10,11].

Despite advances in automated DR classification algorithms, these methods often do not account for the relative importance of different lesions in the grading process and rely solely on a single imaging mode. Fluorescein angiography, another retinal imaging technique, provides additional information about retinal vessels and related lesions with different diagnostic sensitivities in certain regions compared to color fundus images [12]. Combining data from both imaging modalities enhances diagnostic accuracy and efficiency. Due to its invasive nature, fluorescein angiography is unsuitable for routine preventive screening, leading to limited publicly available dual-modal fundus image data. Recent studies have used deep learning-based algorithms to expand this data, improving the performance and robustness of medical image processing algorithms [13]. Synthetic retinal images based on deep learning methods can mitigate imbalances in retinal image samples, providing a data foundation for accurate DR grading using combined color fundus and fluorescein angiography images. However, dual-modal DR grading methods based on color fundus and fluorescein angiography images lack effective feature fusion methods [14,15].

This study aims to design an MVLA-Net multimodal grading model that leverages disease information from different modalities and optimizes the dual-modal feature fusion process. Integrating a CNN encoder with a transformer encoder in MVLA-Net leverages the strengths of both architectures, enabling effective multi-scale feature extraction critical for accurate diabetic retinopathy grading. The dual-branch architecture combines the localized feature extraction capabilities of CNNs with the global context modeling of transformer encoders. This synergy addresses the limitations of single-modal methods, particularly in handling the multi-scale variability of retinal lesions and long-range feature dependencies in fundus images. By integrating these architectures, the proposed model ensures a more robust representation of critical lesion features, resulting in improved accuracy and reliability in diabetic retinopathy grading compared to existing methods. This study has significant scientific implications in theoretical and clinical practice. The proposed vessel segmentation and grading methods quickly and accurately capture retinal vessel information, making them effective tools for assisting clinical DR examinations.

1.1 Deficiencies in Existing Approaches

Diabetic Retinopathy is a prevalent eye condition among diabetic patients and a leading cause of blindness globally. Early diagnosis is crucial for timely treatment and preventing further vision loss. The most common diagnostic method is examining color fundus images, which requires meticulous visual observation by a professional ophthalmologist. Several challenges exist in DR grading diagnosis, including small, easily overlooked lesions, visual similarities in different DR grades, an imbalance in sample sizes across various categories, and complex grading criteria. These factors can lead to misjudgments and classification confusion. However, the scarcity of expert ophthalmologists makes it challenging to meet the needs of many patients with DR. Manual diagnosis is time-consuming and labor-intensive, which can undermine the accuracy of doctors' diagnoses. Therefore, the development of computer-aided diagnostic modeling systems for DR grading has become increasingly important. Currently, mainstream deep learning models are widely used for DR grading diagnosis, but their accuracy is not high due to the complexity of associated lesions in fundus images. This study designs an assisted grading diagnosis model for DR on color fundus images and conducts extensive experiments on relevant public datasets. The results show this model has better diagnostic accuracy and generalization ability than mainstream deep learning models.

1.2 Motivation and Novel Contribution

This study develops retinal vessel segmentation and automated DR grading architecture to enhance the accuracy and reliability of DR diagnostics, thereby forming the basis for a comprehensive system. The proposed model extracts important vascular features for clinical diagnostic assistance. The proposed grading model initially grades color fundus images for DR screening and preliminary diagnosis.

- The study aims to improve DR grading accuracy using the residual structure and upsampling attention from color fundus images without expert intervention.
- The proposed model fuses multi-scale consensus information by coupling local features extracted by a CNN-based encoder with global contextual representations from a Transformer encoder.
- The preprocessing methods were inadequate, leading to the introduction of an interpretable approach using Class Activation Mapping (CAM) to visualize feature extraction areas, localize them using feature maps, and crop abnormal regions of the ophthalmic disease dataset for improved classification accuracy.

The paper's outline follows: Section 2 presents the related studies. Section 3 depicts the data enhancement, feature learning, and proposed framework. Section 4 highlights the study's findings and interprets and evaluates the outcomes. Finally, Section 5 summarizes the study's key findings, novel contributions, and future work direction.

2 Related Work

Accurate diagnosis and treatment of DR depend on precise segmentation and visualization of complex vascular structures. The morphological information from retinal vessels is helpful for early DR diagnosis and assists ophthalmologists with surgical planning and navigation. Over the past decade, DL-based approaches have become prevalent for analyzing fundus images. Accurate DR diagnosis and treatment rely on the segmentation and visualization of complex retinal vascular structures. Neural network decoding layers allow for feature visualization, although they are limited to practical diagnostic use in DR image analysis [16,17]. Thus, precise retinal vessel segmentation remains essential for providing pixel-level visual data for ophthalmologists and clinical research. In fundus image-based segmentation, extracting contextual information and recognizing multi-scale retinal vessels are critical areas of focus [18]. Current research employs global and local relationships within retinal image features to improve the localization of vessels within complex fundus structures and leverage multi-scale features or attention mechanisms to enhance the recognition of various vascular morphologies. CNN-based methods have improved segmentation accuracy by enhancing multi-scale context extraction, whereas boundary information, using edge detectors to capture capillary boundaries, has also proven valuable [19]. Vision Transformer-based methods, which capture global features in fundus images without iterative receptive field expansion, have shown advantages over traditional CNNs. However, their lack of local-global context modeling for multi-scale features limits their application in retinal vessel segmentation. Thus, this study proposes a multi-scale retinal vessel segmentation method combining CNN and Transformer consensus information for feature extraction across scales, allowing adaptive processing of retinal vessels with substantial scale variation.

2.1 Single-Modal DR Grading

Retinal disease diagnosis algorithms focus on rapid and accurate disease classification. While single-modal methods for color fundus or fluorescein angiography images have advanced automatic grading systems, they are limited by incomplete lesion feature information, requiring improved grading accuracy. Hai et al. [20] developed a DRGCNN model to address imbalanced data distribution in DR. The model allocates equal channels to feature maps and introduces a CAM-EfficientNetV2-M encoder for input retinal fundus images. The model incorporates fundus retinal images from both eyes for feature fusion during training. Experimental results show exceptional performance, making it a highly competitive intelligent classification model in DR. Cao et al. [21] proposed WAD-Net methods, which use image block-level annotations and image level to capture fine-grained tumor characteristics for grading, achieving more effective learning through collaborative frameworks. Rodríguez et al. [22] employ transformer models to understand lesion information and color fundus image-based DR grading methods to improve loss constraints and decision interpretability. Due to image scarcity, they rely on private datasets for fluorescein angiography-based grading. Silva et al. [23] developed automated ML-based models to predict DR progression from ultra-widefield retinal images. The proposed approaches were created from baseline on-axis 200° UWF images labeled for progression based on a clinical severity scale. The model was evaluated using a 328-image dataset from the same patient population.

2.2 Multi-Modal DR Grading

Ophthalmologists often employ different imaging techniques for complex cases, as relying on a single examination method is insufficient for identifying all pathological changes. Color fundus images are widely used, capturing intuitive *in vivo* information and features. In contrast, fluorescein angiography is sensitive and accurate in early DR diagnosis, identifying microaneurysms not easily seen in color fundus images. Combining these two modalities offers a broader diagnostic scope, benefiting clinical diagnosis and supporting automated DR detection and severity grading. Current dual-modal DR grading methods combining color fundus and fluorescein angiography images focus on dataset expansion, feature extraction, and fusion strategy selection. Dual-modal methods require effective feature extraction from imaging types and unifying feature representation within a shared space. Given the significant differences in lesion characteristics, specific lesions may be better extracted from one modality. Wavelet transforms can extract the optic disc, vessels, and microaneurysms in fluorescein images, while exudates are detected in color fundus images, and the information is then fused for accurate grading. Bodapati et al. [24] employ pre-trained deep convolutional neural networks to develop an Adaptive Ensemble Classifier for diagnosing diabetic retinopathy. The method achieved 81.86% accuracy on the Kaggle APTOS-2019 benchmark dataset, promising for improving diagnosis in advanced severity grades. Bondala et al. [25] propose Drop Block-based Spatial-Channel Attention U-Net (DB-SCA-UNet), an adaptive system that enhances local features while suppressing irrelevant ones. The method replaces original U-Net convolutional blocks with channel dropout convolutional blocks to mitigate overfitting. Experimental results show that DB-SCA-UNet can accurately and efficiently segment retinal vessels. Bilal et al. [26] have developed an AI-driven VTDR prediction system that integrates multiple models through majority voting. The projected method obtained an accuracy of 99.89%, a sensitivity of 84.40%, and a specificity of 100% on the IDRiD dataset, marking a transformative era in healthcare. This technology ensures prompt and precise VTDR diagnoses, particularly in underserved regions. Due to the scarcity of dual-modal retinal datasets, various image reconstruction methods are used to generate corresponding fluorescein angiography images from color fundus images. The encoder learns lesion representations across modalities during generation, improving downstream lesion segmentation and DR grading tasks. Dual-modal grading in medical images from different modalities faces challenges due to lack of coherence, unrealistic texture, and insufficient high-frequency details. Current methods mainly use pre-trained data but lack comprehensive lesion information during real-time inference, diminishing sensitivity to critical grading information. Morphological changes in retinal vessels are essential for diagnosing DR and indicating early disease symptoms. CNNs have progressed in automated segmentation of retinal vessels, but they have limitations in modeling long-range feature dependencies. Visual Transformers can capture global dependencies but disrupt local spatial details [27]. This study proposes a robust model by integrating transformers and CNNs to address these issues. This model effectively extracts local information and global representations to enable collaborative learning of consensus features. This study leverages lesion information across modalities to address the limitations of single-modality data, optimize feature fusion, and improve the accuracy and reliability of automated DR grading. This method will offer valuable scientific insights for theoretical and clinical applications.

3 Methodology

Research on retinal disease diagnosis algorithms aids in identifying and localizing retinal pathological changes, providing diagnostic tools for ophthalmologists, collecting ophthalmic data, and supporting clinical research through statistical analysis. Rapid and accurate classification of retinal diseases, assisting doctors with potential diagnostic inferences and treatment suggestions, enhancing prevention, diagnosis, and follow-up services, and promoting large-scale eye disease screening. Also, advanced augmentation techniques, such as vessel-focused transformations and intensity scaling, were applied to enhance dataset diversity and robustness. To assess the efficacy of the hierarchical fusion network model, both quantitative and qualitative experiments were conducted on the color fundus images. The model's performance was evaluated using various metrics to assess image quality. A qualitative analysis compared natural and synthetic DR images, focusing on crucial structures like blood vessels and microaneurysms to assess its performance in generating critical features.

3.1 Experimental Dataset

The success of deep learning in DR grading is closely tied to the datasets used. Unlike other discrete classification tasks with no apparent progression, typical DR grading settings often disregard the continuity between different DR severity levels, treating the grades as distinct and separable categories. The proposed models are accessed using fundus images taken from public datasets such as EyePACS, IDRID, and multi-APTOS 2019 datasets, and the quality of the datasets and the accuracy of the annotation directly affect the DR classification.

EyePACS dataset is a large-scale and high-quality dataset widely used to evaluate the performance of DR classification tasks [28]. Many images have issues with resolution, such as blurriness, defocusing, and exposure problems, due to differences in equipment and environmental conditions across collection sites. The EyePACS dataset labels DR into five levels, from DR0 to DR4, representing a range from no lesion to severe lesion, containing 970 training and 243 test samples, totaling 1213 images with varying resolutions ranging from 433 × 289 to 5184 × 3456 pixels.

The **IDRID dataset** [29] is a public dataset that includes 516 fundus images of 4288 × 2848 pixels resolution and in JPG format. In the IDRID dataset, 80% of the images (413) form the training set, while the remaining 20% (103) constitute the test set. Medical experts graded all images for both DR and DME severity.

The **Multi-APTOS 2019 public dataset** [30] consists of 3662 color fundus instances, with the majority being 3216 × 2136 pixels. The training set consists of 80% of the images from each class, while the testing set includes the remaining 20%. The image distribution of three datasets into training (80%) and test (20%) sets, is illustrated in Table 1.

Table 1: Data split settings for EyePACS, IDRID, and multi-APTOS 2019 datasets

Dataset	Total images	Training set (80%)	Test set (20%)
EyePACS	1213	970	243
IDRID	516	413	103
Multi-APTOS 2019	3662	2929	733
Total	5391	4312	1079

3.2 *Image Enhancement*

In image processing, selecting and adjusting color channels is a common approach. For fundus images, the green light wavelength penetrates the retina more effectively, making vascular and structural details more pronounced. Extracting the grayscale image of the green channel from the RGB image enhances contrast by reducing redundant features. This study proposes combining the green channel grayscale image with Ben's enhancement to improve classification performance. Cropping these areas in preprocessing can reduce redundant information and lighten the model's computational load. Standardizing all images to a fixed resolution prepares them for feature extraction. This study used Ben Graham, a novel image enhancement technique for a diabetic retinopathy competition. This technique involves Gaussian blur to reduce noise and smooth details, yielding significantly improved classification results.

3.3 *Multi-Modal Data Augmentation*

Traditional data augmentation techniques in DR image classification networks include removing black borders, adjusting resolution, normalization, and geometric transformations. This study employed advanced augmentation strategies, including geometric transformations, noise injection, domain-specific enhancements, and intensity standardization, to simulate diverse imaging conditions, enhancing model robustness and performance. Geometric transformations increase model robustness by diversifying data, while traditional methods fail to introduce essential information in multimodal retinal disease classification tasks. Due to the scarcity of multimodal retinal datasets, GANs have been widely applied in multimodal retinal data analysis models. This study employed GAN-based image generation techniques due to their ability to address imbalances in retinal image distribution and patient privacy concerns.

3.4 Dual-Modal Fundus Image Registration and ROI Extraction

The fundus images use a pixel-level correspondence with a specific domain approach that aligns features across modalities, relying on retinal vessels in proposed imaging modalities [14,31]. The registration process involves a weakly supervised deformable registration method implemented on the dataset. Initially, color fundus images are input to two unsupervised segmentation networks, obtaining vascular segmentation results for each. Intermediate features are extracted from these networks and fed sinto a registration estimation network, predicting the registration field based on vascular features. However, regions of interest (ROIs) in each image do not entirely overlap, and each image's ROI is extracted separately. Following dual-modal retinal image registration and overlapping ROI extraction, each dual-modal image achieves complete feature alignment within the defined area, providing a robust foundation for modality transformation between paired images. To evaluate the effectiveness of MVLA-Net in handling multi-scale variations, vessel thicknesses were categorized as thin (≤2 pixels), medium (3–7 pixels), and thick (>7 pixels). Metrics such as DSC and QWK were calculated for each category, demonstrating robust performance across all scales, particularly for thin vessels, due to the inclusion of a SAM.

3.5 Proposed Models for Ophthalmic Diseases

Attention-based approaches are commonly used in diabetic retinopathy grading to extract valuable lesion feature information. Typical attention mechanisms, including channel and spatial attention, are integrated with CNN architecture. However, these methods often pose class imbalance challenges and difficulty classifying small sample categories. CNN-based methods are limited in capturing long-distance relationships due to their limited receptive field. This study uses fundus images to introduce a multi-view lesion attention network (MVLA-Net) for DR grading. MVLA-Net integrates pathological information from images and presents a dual-modal lesion feature interaction module, modeling lesion relationships using semantic information from both modalities. This allows more accurate and comprehensive capture of lesion features in fundus images, resulting in superior DR grading performance compared to traditional methods. To evaluate the effectiveness, this study compares the ophthalmic diseases grading model with cutting-edge transfer learning-based CNN architectures, including mVGGNet19, InceptionV3, ResNet120V2, DenseNet121, EfficientNetB7, and ConvNext. Fine-tuning was performed by adjusting classifier parameters and modifying the number of categories in the final layer to meet the classification needs of the smart diagnostic model for ophthalmic diseases.

3.5.1 Proposed Multi-View Lesion Attention Network (MVLA-Net) Architecture

This study proposes the MVLA-Net architecture to optimize the DR prognostics by employing a multi-view attention block (MAB) and a dual-modal feature interaction module (DFIM), guided by the Spatial Attention Module (SAM) for precise dual-model integration, as shown in Fig. 1. MVLA-Net enables to learn robust feature representations, focusing on decision-making based on valuable grading information. The dual-encoder design in MVLA-Net combines the CNN's ability to extract fine-grained spatial features with the transformer's capacity to model global contextual information. This integration addresses the limitations of single-modality encoders, enabling the network to process complex retinal vessel morphologies and lesion structures effectively. The model processes fundus images as inputs, fusing extracted features through channel concatenation to optimize representation. However, the class imbalance remains challenging, particularly affecting recognition accuracy for underrepresented classes. To address this, MVLA-Net integrates a multi-view attention module post-feature extraction to process fused dual-modal features, enhancing intra-class information capture, enhancing differentiation, and optimizing category semantics within the dual-modal grading framework. The DFIM focuses on lesion features integration, emphasizing shared lesion characteristics while compensating for missing details. This integration generates lesion heatmaps that guide the SAM for accurate lesion localization. The SAM output is passed through a classifier with global average pooling and a fully connected layer for severity grading. MVLA-Net offers enhanced clinical relevance by providing grading decisions and improved feature extraction by combining MAB and SAM guided by DFIM.

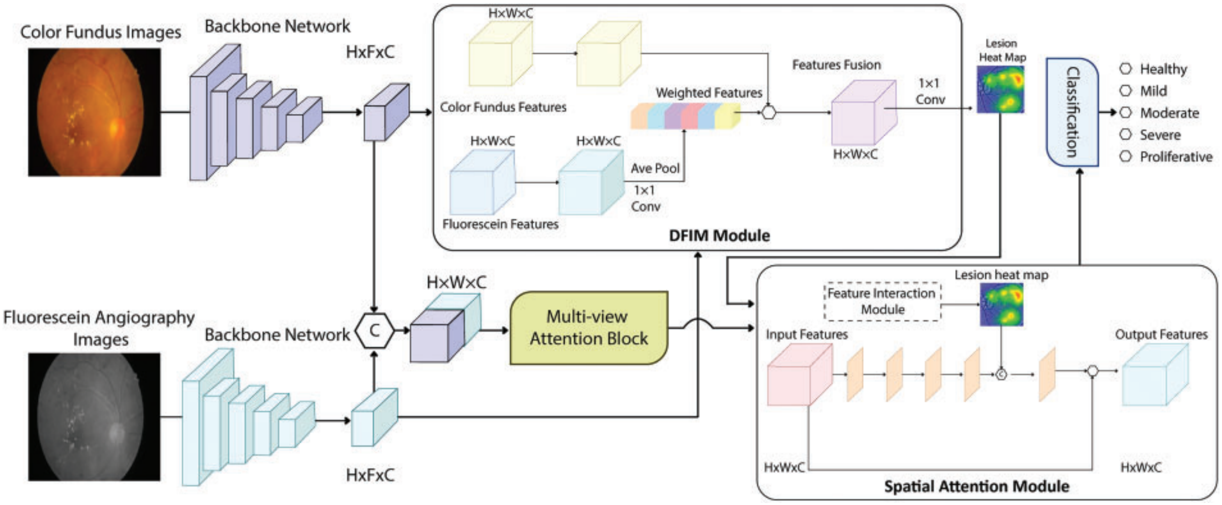


Figure 1: Overview of the MVLA-Net architecture, showing the integration of CNN and transformer encoders, attention modules, and classification pipeline for diabetic retinopathy grading

3.5.2 Multi-View Attention Block (MAB)

Because of the specific nature of the medical analysis, most diabetic retinopathy datasets face significant challenges with class imbalance. Certain classes contain only a few samples, limiting the model's ability to learn effectively from these small datasets and significantly impacting recognition accuracy for these underrepresented classes. This imbalance reduces DR algorithms' classification performance and generalization capability, making it an urgent issue. Additionally, the symptom similarity across adjacent DR severity levels complicates accurate model predictions. This study designs a *multi-view attention* module to address class imbalance and high symptom similarity. The module uses a 1 × 1 conv layer to create a feature fusion. $Y \in R^{H\times W\times k}$ where L is category measures, and k is a preset number of channels allocated to extract features for each category. A channel split operation divides Y into L groups, enabling parallel processing of intra-class features for each category. Dilated convolution is applied to each y_i' to increase each class's receptive field and reduce inter-class information interference. Global average pooling and channel-wise average pooling along the category dimension are applied to obtain the attention weight for each category. This boosts the algorithm's ability to distinguish between distinct severity levels.

3.5.3 Spatial Attention Module (SAM)

MVLA-Net is an innovative imaging system employing color fundus images to accurately model key pathological regions for early-stage DR grading. The farmwork leverages spatial attention guided by modality interaction information to capture critical lesions. The system's primary architecture consists of DFIM and SAM structures, with a routing algorithm in capsule attention to estimate internal components and transformer-guided spatial attention to capture long-distance dependencies for precise localization.

The DFIM module establishes dependencies between modalities, focusing on shared critical lesions and extracting contextual correlations. This module employs an attention-weighted fusion process based on critical lesion locations captured by DFIM, achieving accurate lesion localization. The primary feature, F_{CF}, is extracted from color fundus images, while complementary features, F_{FFA} generate an interaction (ICT) weighted feature W_1, as shown in Eqs. (1) and (2).

$$F_{\mathrm{ICT}} = \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}_{3\times3}(F_{FFA}))) \tag{1}$$

$$W_1 = \mathrm{sigmoid}\left(\mathrm{Conv}_{1\times1}\left(\mathrm{GAP}\left(F_{\mathrm{ICT}}\right)\right)\right) \tag{2}$$

Interaction-weighted feature W_1 is used to adjust the importance of each channel in the primary feature F_{CF}, and the adjusted feature map $F_{\mathrm{interaction}}(F_{\mathrm{ICT}})'$ generates a lesion heatmap M for guiding spatial attention in localizing key lesions, as indicated in Eqs. (3) and (4).

$$F'_{F_{\mathrm{ICT}}} = F_{CF} \otimes W_1 \tag{3}$$

$$M = \mathrm{Conv}_{1\times1}(FF_{\mathrm{ICT}}{}'_{.}) \tag{4}$$

Spatial attention captures spatial correlations among features, focusing on critical lesion areas while suppressing irrelevant regions. The average pooled along with channel dimension is applied in the spatial attention branch to the multi-attention optimized dual-modal fusion feature F_{multi}, producing spatial feature descriptor W_2 through sigmoid activation, as represented in Eq. (5).

$$W_2 = \mathrm{sigmoid}(\mathrm{CGAP}(F_{multi})) \tag{5}$$

The lesion heatmap M from DFIM refines the spatial descriptor W_2, resulting in the final spatial score W_3, and the spatial attention output F_{spatial}, as shown in Eqs. (6) and (7).

$$W_3 = M \otimes W_2 \tag{6}$$

$$F_{\mathrm{spatial}} = F_{multi} \otimes W_3 \tag{7}$$

By guiding spatial attention with the lesion heatmap M, MVLA-Net can identify critical lesion features, allowing for more accurate localization of crucial lesions. Finally, F_{spatial} is passed through global average pooling and a fully connected layer to predict the severity level of diabetic retinopathy.

3.5.4 Analysis of Multi-View Lesion Attention and Data Fusion Methods

This study explores multimodal data fusion techniques for multimodal fundus imaging, focusing on the automatic detection, association, and combination of data from multiple sources. The primary motivation is to harness the unique characteristics of multi-modalities to create a unified representation, facilitating improved decision-making. Multimodal data fusion methods can be categorized into early, late, and hybrid. Early fusion generates a common feature vector for specific processes like classification or segmentation. Late fusion allows interaction and enrichment of lesion information between different modalities. This work explores feature-level fusion to improve model capacity for extracting class-relevant information from related modalities, compensating for single-modality imaging deficiencies and providing a more precise classification of diabetic retinopathy. Late fusion combines independent decisions of each modality to arrive at the last decision. Hybrid fusion integrates early and late fusion mechanisms, enhancing diversity and flexibility but complicating training. The performance of hybrid methods depends on selecting the integration method, known as the combination strategy.

4 Experimental Results Analysis and Discussion

This study evaluates the efficacy of MVLA-Net in diabetic retinopathy grading using three public datasets. Ablation experiments demonstrate the advantages of the dual-modal grading approach over single-modal methods. The analysis discusses how different fusion strategies and parameter choices impact the model's efficacy. The dual-branch design of MVLA-Net outperforms traditional approaches by significantly improving accuracy and feature extraction efficiency. The transformer encoder captures the long-range dependencies among lesion features, while the CNN focuses on fine-grained spatial features. This combination enhances lesion detection and classification, as reflected in the performance metrics where the model surpasses other state-of-the-art architectures.

4.1 Experimental Environment Setup

Using the PyTorch framework, the proposed MVLA-Net model trains on NVIDIA GTX 4050Ti GPU. The training process employed a stochastic gradient descent (SGD) optimizer with a momentum parameter 0.9 and a weight decay 0.0005. The training was conducted for 60 epochs, 100 steps per epoch. The first 20 epochs were used as a warm-up phase, during which the learning rate gradually increased from 0 to a maximum of 0.001, after which it followed the cosine decay schedule, and the learning rate progressively decreased to 0.00005. Batch size 32 was applied, with 512 × 512 resolution color fundus as input to the model. No additional data augmentation strategies were applied during training to verify the advantages of dual-modal data incorporating generated images over single-modal data. The study aimed to optimize the performance of the muti-view Attention block in a dual-modal grading model by analyzing the effects of fixed channel number k and dilation rate d on model performance.

4.2 Evaluation Metrics

A confusion matrix typically represents a model's prediction results in classification tasks. Researchers can calculate and assess various model capabilities based on the differences between predicted and actual values in the confusion matrix.

Accuracy represents the proportion of accurately predicted instances to the total number of samples, providing an overall measure of the model's classification performance. It is calculated as in Eq. (8).

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

where TP and TN are the counts of true positive and true negative classifications, respectively, and FP and FN represent the counts of false positive and false negative classifications, respectively.

Precision is the proportion of true positives among the sample the model predicted as positive. It measures the accuracy of the model's positive predictions and is calculated as in Eq. (9).

$$PRE = \frac{TP}{TP + FP} \tag{9}$$

Recall is the percentage of actual positive samples correctly predicted by a model, focusing on its coverage of all true positives, and is calculated as in Eq. (10).

$$REC = \frac{TP}{TP + FN} \tag{10}$$

Specificity measures the percentage of negative samples correctly predicted as favorable, evaluating the model's accuracy in negative predictions, calculated as in Eq. (11).

$$Specificity = \frac{TN}{TN + FP} \tag{11}$$

FScore is the harmonic mean of Precision and Recall, considering accuracy and coverage. It is particularly suitable for imbalanced datasets, and its formula is depicted in Eq. (12).

$$FScore = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{12}$$

AUC represents the area under the ROC curve, where the ROC curve's x-axis and y-axis represent the False Positive Rate (FPR) and True Positive Rate (TPR), respectively, calculated as in Eqs. (13) and (14).

$$TPR = \frac{TP}{TP + FN} \tag{13}$$

$$FPR = \frac{FP}{FP + TN} \tag{14}$$

AUC values range between 0 and 1, with higher values indicating better model classification performance, particularly in scenarios with imbalanced data.

Quadratic Weighted Kappa (QWK) measures the agreement between the model and actual observations, accounting for consistency across categories, making it especially useful for multiclass classification tasks.

$$QWK = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}} \tag{15}$$

It is calculated as in Eq. (15), where $O_{i,j}$ is the actual likelihood, $E_{i,j}$ is predicted likelihood, N is the total number of categories and i and j represent any two categories in the set.

4.3 Result Analysis Based on Accuracy

This study conducted experiments based on three datasets, APTOS 2019, EyePACS, and IDRiD, which diabetic retinopathy categorizes into five categories given the clinical necessity of screening for low-quality images. This study employs the AUC and accuracy as performance evaluation metrics to assess a classifier's performance. Although many other metrics exist, this study focuses on AUC and ACC to simplify and standardize the evaluation. The study uses the proposed model and other CNN-based architectures to explore the impact of DFIM and attention mechanisms on DR classification. It was found that adding DFIM and attention modules improved DR accuracy and precision but slightly decreased recall. However, when DFIM and attention modules were added sequentially to the proposed model, DR classification accuracy and AUC reached their highest level. The study underscores the importance of considering DFIM and attention mechanisms' influence on DR classification performance. The DFIM designed pathological information and guided spatial attention for precise lesion localization.

Comparative experiments were performed to assess the impact of DFIM on lesion localization using color fundus features. In DFIM, color fundus features drive lesion localization, which is potentially overlooked in color fundus imaging. This configuration achieves superior grading performance, with an ACC of 98.4% and a QWK of 0.981, as depicted in the training and validation accuracy in Fig. 2. The model demonstrates a steady improvement in training and validation accuracy throughout 60 epochs, stabilizing close to peak performance and indicating effective generalization and robustness in the model's learning process. Including DFIM for lesion localization improves overall model performance, yielding a 0.3% higher ACC than configurations without feature fusion. The loss curves also reflect consistent convergence with reduced overfitting, as evidenced by the close alignment of training and validation losses. Fig. 2 reveals a unique pattern where validation loss is lower than training loss due to regularization mechanisms like dropout and weight decay. The validation set often contains simpler examples, resulting in lower validation loss during early training stages. The learning rate schedule, which includes a warm-up phase, slows training loss convergence while maintaining stable validation loss. The training and validation loss curves converge as training progresses, demonstrating the model's generalization ability across datasets. The AUC metric was used to evaluate the DR classification experiment. A quadratic weighted kappa metric was applied for the five DR classification experiments, effectively reflecting model performance on imbalanced datasets. The kappa value ranges from 0 to 1, with higher values indicating better model performance. Additionally, the ROC curve in the figure compares the proposed MVLA-Net with baseline models, highlighting its competitive performance with an AUC of 0.99.

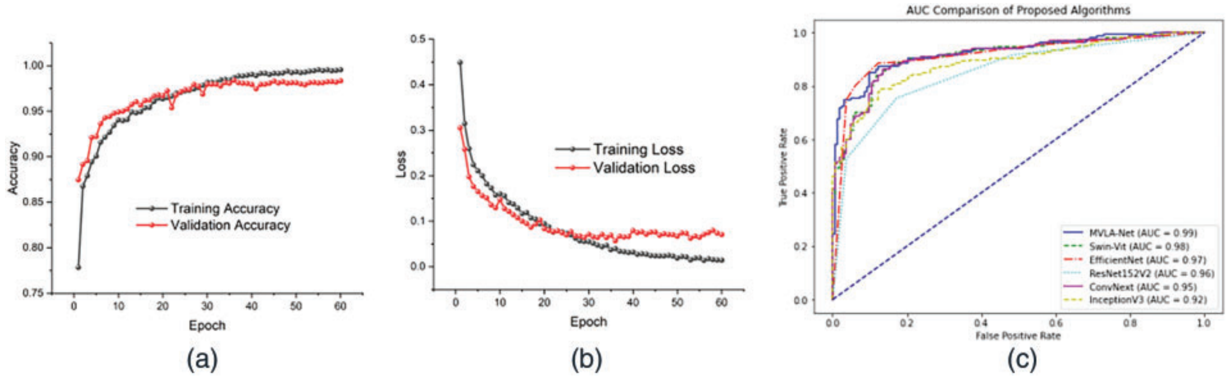
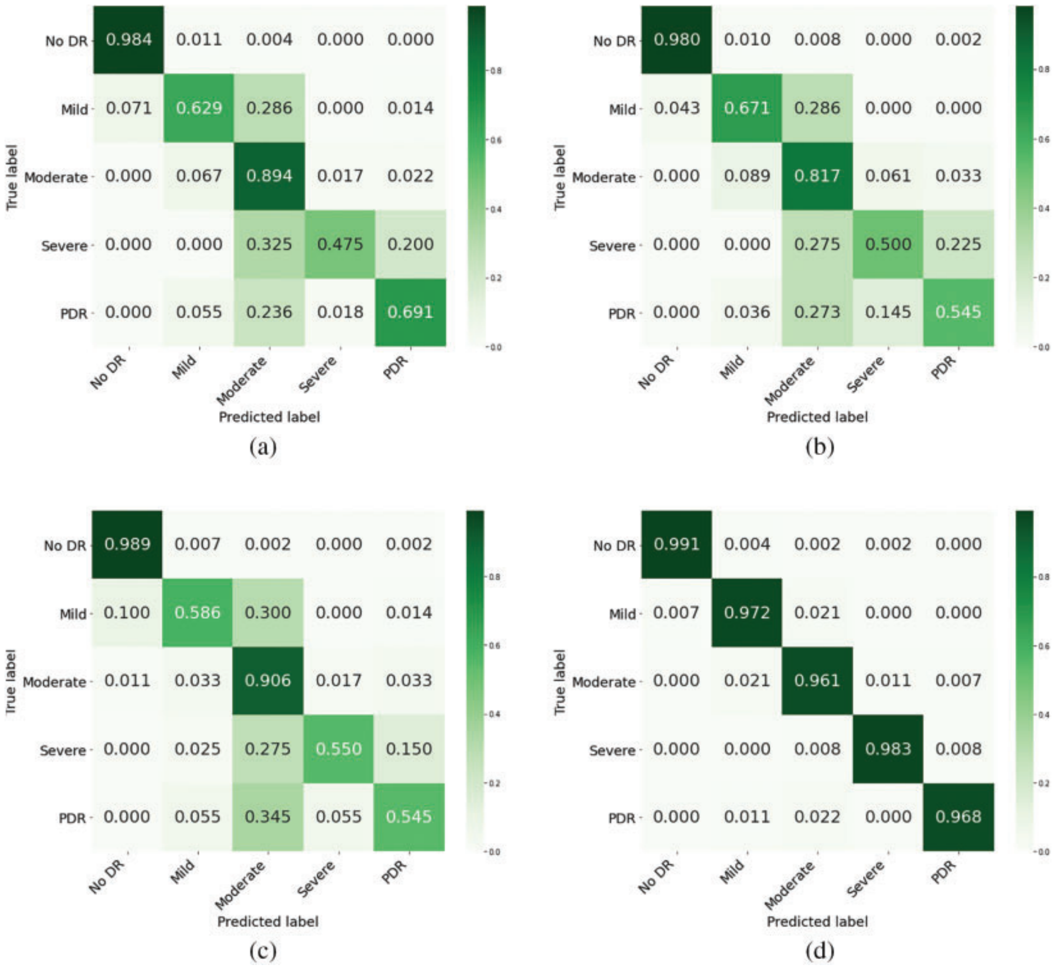


Figure 2: Depicts the performance of the proposed models, (a) represents the model's accuracy, (b) loss of the proposed model, and (c) AUC comparison of the proposed with the SOTA models

4.4 Comparison Based on Confusion Matrix

The confusion matrix is a tool used to evaluate a classification model's performance on a test set. It shows the actual classes and predicted classes. This study analyzed the confusion matrices of proposed models and CNN-based architectures. The results showed that MVLA-Net achieved more accurate predictions than CAB-net, ConvNext, Efficient Net, and others, as shown in Fig. 3. The improvement of MVLA-Net over CAB-net indicates the successful addition of a differential regression task. The DFIM and attention module were also influential in enhancing feature extraction. The proposed MVLA-Net model is more effective for predicting the severity of DR. Although the model may fail in some cases, it brings predicted labels closer to actual labels, indicating better extraction of significant features and patterns from the training data. The study reveals that decision-level fusion models generally have lower grading performance, but feature-level fusion consistently enhances it across all model configurations. This is due to the higher information richness in feature-level fusion, which combines color fundus images into a unified feature vector. This enables the model to learn better interactions and relationships between modalities. The MAB and modality interaction-guided spatial attention modules are better suited for optimizing fused features, mitigating class imbalance issues, and identifying distinguishing features between classes.



(a) (b) (c) (d)

Figure 3: Confusion matrix of the proposed model compared to the top three models with the highest accuracy, (a) representing CABNet, (b) ResNext, (c) Efficient Net, (d) proposed model

4.5 Comparison with Other Approaches

To evaluate the DR grading performance of the MVLA-Net model based on DFIM and attention module in the dual-modal grading model is compared to other models that reported results on three datasets. Table 2 depicts the hybrid CNN-transformer architecture in MVLA-Net achieves superior performance across all evaluation metrics compared to CNN-only and transformer-only architectures. This demonstrates the efficacy of combining local and global feature extraction mechanisms. Experimental results show that these models are unsuitable for diabetic retinopathy grading. The proposed approach demonstrates a distinct advantage in locating critical lesion information within dual-modal fusion. The proposed model achieves the highest accuracy, recall, and precision compared to other models. It also performs best on the FScore, which balances recall and precision. It can be seen that the multi-view lesion grading network achieved strong performance in DR grading, with AUC, Recall, and FScores of 0.992%, 98.2%, and 98%, respectively, on the APTOS dataset. These results ranked first among the compared methods, surpassing the best results from other methods by 0.6%, 0.7%, and 0.6%, respectively. While aimed at mitigating overfitting and utilizing prior knowledge, pre-trained basic deep learning models still performed poorly, with classification results among the lowest. However, DenseNet showed unexpectedly low recall, likely due to DenseNet's highly converged output, which may have discarded too much graph structure information, thereby limiting spatial information capture. The high parameter counts in DenseNet, combined with small batch sizes, may have restricted its ability to learn sufficiently. Tables 2 and 3 show the experimental results for the validation and test sets on the three datasets, respectively. The empirical findings for mVGGNet19, InceptionV3, ResNet152V2, DenseNet121, Efficient Net, ConvNext, and MVLA-Net were taken accuracy, Kappa, and other metrics. It can be observed that MVLA-Net, the method proposed, demonstrated a significant advantage in classification tasks over other mainstream CNNs. MVLA-Net achieved the highest accuracy and Kappa scores for the validation and test sets. Specifically, MVLA-Net ACC was 99% for the validation set, and Kappa was 98%, outperforming the best results from other networks by 3.7% and 5.8% on the APTOS 2019 dataset, respectively. It can be observed that MVLA-Net outperformed DP2M-Net [32] in ACC, and Kappa, on the validation set, exceeded DP2M-Net results by 1.9% and 3.6%, respectively. In Table 2, the performance of MVLA-Net is comprehensively compared against state-of-the-art models across the APTOS 2019, EyePACS, and IDRiD datasets. MVLA-Net consistently achieved superior metrics, including accuracy, precision, recall, F1-score, AUC, and QWK. Specifically, MVLA-Net achieved the highest accuracy of 98.43% on APTOS 2019, surpassing the next-best model by 0.7%. The dual-modal architecture's capability to combine global and local feature representations is pivotal in these results.

Table 2: Performance comparison with other cutting-edge approaches

Model	APTOS 2019 dataset performance in (%)						EyePACS dataset performance in (%)						IDRiD dataset performance in (%)					
	ACC	REC	PRE	AUC	FS	QWK	ACC	REC	PRE	AUC	FS	QWK	ACC	REC	PRE	AUC	FS	QWK
mVGGNet19	90.97	91.54	92.32	0.915	0.917	0.915	92.72	91.84	92.17	0.936	0.923	0.936	82.2	86.5	87.1	0.849	86.7	0.849
InceptionV3	93.97	94.45	92.89	0.933	0.952	0.933	92.52	94.73	93.78	0.936	0.945	0.936	87.8	87.3	89.2	0.873	86.9	0.873
ResNet152V2	93.43	92.29	91.02	0.921	0.916	0.921	91.52	91.17	92.64	0.903	0.914	0.903	93.1	92.6	93.5	0.921	93.3	0.921
DenseNet121	94.84	93.46	94.02	0.934	0.946	0.934	93.33	92.84	92.64	0.931	0.942	0.931	86.5	87.8	86.8	0.867	87.3	0.867
EfficientNet	96.98	97.37	98.22	0.973	0.954	0.973	96.77	95.51	96.32	0.958	0.964	0.958	93.8	92.4	91.9	0.924	92.6	0.924
ConvNext	94.87	95.83	97.41	0.957	0.947	0.957	97.03	94.22	94.89	0.953	0.962	0.953	92.8	91.3	90.8	0.912	91.8	0.912
Swin-ViT	97.25	96.89	97.53	0.982	0.978	0.978	97.53	96.87	97.42	0.979	0.975	0.976	95.1	94.6	94.9	0.968	97.4	0.968
CoVi-Net	96.70	96.12	96.85	0.975	0.970	0.972	96.12	95.88	96.33	0.971	0.968	.970	94.2	93.8	94.0	0.952	93.9	0.953
MVLA-Net	98.43	98.20	98.43	0.992	0.983	0.981	99.81	98.05	98.77	0.981	0.986	0.979	98.1	0.981	0.985	0.988	0.989	0.978

Table 3: Comparison of outcomes with cutting-edge research

Algorithms/Approaches	APTOS2019		IDRiD		EyePACS		Parameter
	ACC	QWK	ACC	QWK	ACC	QWK	
ResNet152V2 [33]	0.934	0.921	0.621	0.696	0.915	0.903	58.3 M
ConvNext [34]	0.948	0.957	0.611	0.682	0.97	0.953	49.5 M
Efficient Net [35]	0.969	0.973	0.634	0.698	0.967	0.958	66.1 M
CABNet [36]	0.883	0.925	0.609	0.691	0.811	0.811	44.3 M
mVGGNet19 [37]	0.909	0.915	0.602	0.682	0.927	0.936	65.2 M
Swin-VIT [27]	0.845	0.873	0.612	0.674	0.789	0.761	88.0 M
HA-Net [38]	0.855	–	0.664	–	0.804	0.792	61.3 M
QTL-DR [39]	0.834	–	0.619	–	0.781	0.763	42.7 M
GF-CapsNet [40]	0.865	–	0.641	–	0.801	0.784	47.6 M
MAPCRCI-DMPLC [12]	0.989	0.985	0.695	0.740	0.995	0.985	40.5 M
DP2M-Net [32]	0.989	0.994	0.698	0.742	0.9996	0.986	41.8 M
PathSeg-MorphNet [41]	0.965	0.957	0.675	0.715	0.940	0.920	39.2 M
Proposed MVLA-Net	0.984	0.981	0.981	0.978	0.998	0.979	41.1 M

4.6 Heatmap Visualization

This study uses Gradient-weighted Class Activation Mapping (Grad-CAM) to extract key features for classifying diabetic retinopathy images. Grad-CAM quantifies the activation map, indicating the impact of different parts of the input image on the model's final classification. The model's steps involve obtaining the feature map, calculating the gradient for each class, and backpropagating the gradients to weight and sum the feature map. The output heatmap shows that the model focuses more on lesions like hard and soft exudates and microaneurysms, covering most lesion areas with few missed or incorrectly focused regions. This confirms the proposed model's feasibility and enhancement method for diabetic retinopathy classification, as shown in Fig. 4. These visualizations indicate that MVLA-Net more accurately localizes critical lesion features, such as microaneurysms and exudates, than EfficientNet and ConvNext. The robust localization further reinforces the reliability of the proposed model in clinical scenarios.

4.7 Ablation Study on the Dual-Modal Feature Interaction Module

The study evaluated the impact of every component within MVLA-Net on diabetic retinopathy classifying efficacy through ablation experiments. The results showed that MVLA-Net, with all components included, achieved the best grading accuracy and category consistency performance. The study used single-modal baselines, dual-modal baselines, MAB, and a modality-specific attention module. The results depicted the importance of considering modality interaction information in grading. Initial results show that compared to CABNet, ResNext, Efficient Net, and other models, the proposed model shows a substantial improvement in grading accuracy. Despite the additional data augmentation, this improvement indicates that incorporating lesion information significantly enhances the model's grading performance, highlighting the advantages of a dual-modal grading approach. To further mitigate class imbalance and improve the model's ability to distinguish between adjacent classes, the MAB was added to the Dual-DenseNet121 baseline. Ablation studies demonstrate the effectiveness of the applied augmentation techniques, improved the model's accuracy by 4%, and reduced overfitting, as evidenced by the lower validation loss curve. This addition, which focuses on multi-level attention to the fused features, increased grading accuracy from 96.5% to 98.8% on three proposed datasets. The proposed MVLA-Net incorporates the modality-specific attention module and optimizes the feature representation obtained from MAB, raising ACC and QWK by 5% and 5%, respectively. Ablation studies revealed that incorporating synthetic dual-modal data improved grading accuracy by 3% and lesion localization by 2.5% compared to models trained solely on real single-modal datasets. This improvement is attributed to the attention module's ability to capture lesion relationships across different modalities, locate lesions, and enhance the dual-modal grading model's accuracy and consistency. Table 3 shows that MVLA-Net achieves robust segmentation across all vessel thickness categories and accurately segments thin capillaries and smooth boundaries for thick vessels, highlighting the model's robustness in addressing extreme multi-scale variations.

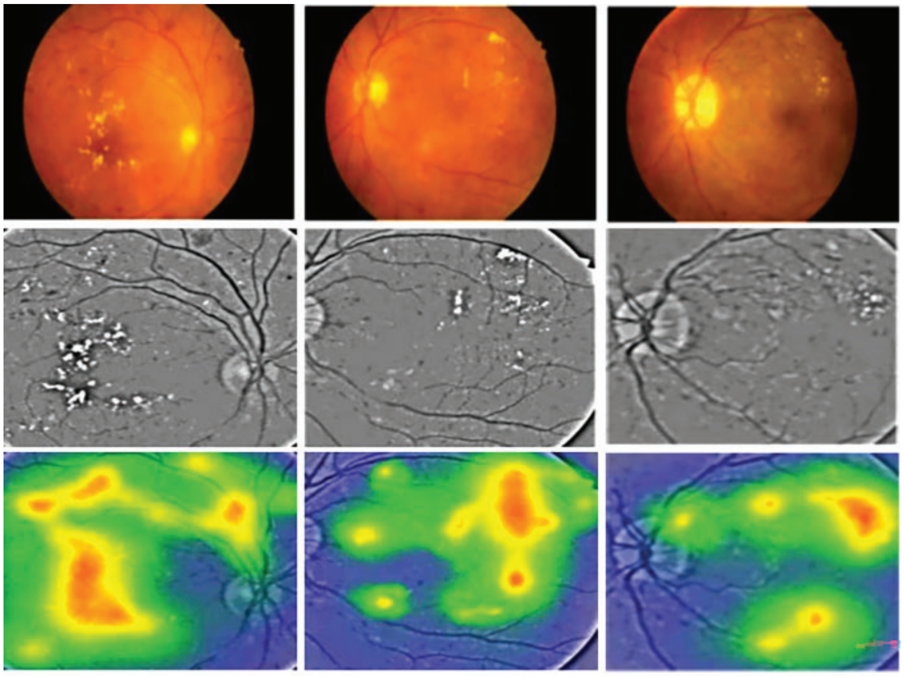


Figure 4: Classification heat map of proposed method

4.8 Comparison with SOTA Methods

This study evaluated the MVLA-Net model against other state-of-the-art models and methods for five DR grading categories. The results showed that the MVLA-Net model achieved the best results across all five categories. ResNet152V2 and EfficientNet have strong feature extraction capabilities but need help accurately classifying specific DR stages due to the complexity and specificity of the task. ConvNext, a recent advanced CNN model, performs comparable to transformer-based models but has limited feature extraction capacity due to depth-wise convolution. GF-CapsNet, DP2M-Net, PathSeg-MorphNet, and CABNet, which incorporate various attention mechanisms, still face limitations in recognizing specific lesions within each DR category. The MVLA-Net has an advantage in differentiating specific categories, eliminating interference from unrelated factors, and reducing the overall difficulty of grading. The MVLA-Net model achieves high grading accuracy and class consistency but needs improvement without supplemental information. It captures critical lesion features through semantic information interaction, allowing more effective extraction of lesion relationship information. Its high QWK metric ensures the model's reliability, which provides robust consistency across grading categories. In this study, MVLA-Net achieved a high QWK of 0.992, demonstrating its reliability in grading accuracy and classification consistency.

Ablation studies and comparative experiments demonstrate MVLA-Net's ability to outperform existing models like DenseNet, ResNet152V2, and ConvNext in handling imbalanced datasets and multi-scale lesion detection. As depicted in Fig. 3, the confusion matrix underscores MVLA-Net's lower misclassification rate, especially for challenging DR categories, validating its enhanced feature extraction capabilities.

5 Conclusion and Future Work

DR is a common ophthalmic disease worldwide, posing a risk of irreversible blindness in severe cases. Due to the scarcity of specialized ophthalmologists, developing a computer-aided diagnostic system for DR grading has become increasingly important. However, current mainstream deep learning methods still struggle to grade DR severity accurately, and their unreliable results limit clinical utility. This study addresses this gap by designing an innovative auxiliary grading model for diabetic retinopathy in color fundus images. By analyzing the characteristics of DR across five distinct stages, we frame the DR diagnostic task as a five-class classification problem in color fundus images. The study introduces the MVLA-Net to improve diabetic retinopathy grading accuracy. The proposed integration of CNN and transformer encoders in MVLA-Net addresses the limitations of single-modality feature extraction, providing a robust framework for diabetic retinopathy grading. It integrates multi-view Attention and Modality Interaction-Guided Spatial Attention to optimize the fusion process of dual-modal information. Ablation studies show that MVLA-Net outperforms single-modal grading methods in diabetic retinopathy tasks without traditional data augmentation techniques. The proposed MVLA-Net demonstrated robust performance across vessel scales, accurately segmenting thin capillaries and thick retinal arteries. Future research could simplify CNN models, reduce parameters, and promote AI diagnostic tools in primary care settings to increase ophthalmic disease detection rates. The MVLA-Net model is a highly accurate and reliable tool for grading diabetic retinopathy in retinal images. It could also be a decision support tool, providing ophthalmologists with additional insights into retinal pathologies. MVLA-Net's high accuracy and processing ability make it suitable for large-scale screening programs, automating initial screenings and prioritizing cases requiring detailed evaluation. Its consistent performance can reduce interobserver variability, improving patient care quality. Furthermore, MVLA-Net can be integrated into telemedicine platforms, enabling remote diagnosis and grading of diabetic retinopathy and expanding access to quality eye care in underserved or rural areas.

Acknowledgement: This research was supported by Princess Nourah bint Abdulrahman University and Researchers Supporting Project. The authors would also like to thank Prince Sultan University, Riyadh Saudi Arabia for support of APC.

Funding Statement: This research was funded by Princess Nourah bint Abdulrahman University and Researchers Supporting Project number (PNURSP2025R346), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Author Contributions: Tariq Mahmood, Tanzila Saba, Faten S. Alamri, and Alishba Tahir conceived and experimented. Tariq Mahmood and Tanzila Saba designed the methodology. Tanzila Saba and Faten S. Alamri reviewed and revised the study. Tanzila Saba, Alishba Tahir, and Faten S. Alamri provided essential research resources. Tariq Mahmood, Alishba Tahir, and Noor Ayesha contributed to the analyzed data. Tanzila Saba, Faten S. Alamri, and Noor Ayesha conducted the proofreading of the study. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The authors confirm that the data supporting the findings of this study are available within the article in Section 3.1.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Jian M, Chen H, Tao C, Li X, Wang G. Triple-DRNet: a triple-cascade convolution neural network for diabetic retinopathy grading using fundus images. Comput Biol Med. 2023;155(9):106631. doi:10.1016/j.compbiomed.2023.106631.
2. Huang Y, Lyu J, Cheng P, Tam R, Tang X. SSiT: saliency-guided self-supervised image transformer for diabetic retinopathy grading. IEEE J Biomed Health Inform. 2024;28(5):2806–17. doi:10.1109/JBHI.2024.3362878.
3. Manarvi IA, Matta NM. Investigating information needs of Saudi diabetic patients. Curr Diabetes Rev. 2019;15(2):149–57. doi:10.2174/1573399814666180612080718.
4. Shoaib MR, Emara HM, Zhao J, El-Shafai W, Soliman NF, Mubarak AS, et al. Deep learning innovations in diagnosing diabetic retinopathy: the potential of transfer learning and the DiaCNN model. Comput Biol Med. 2024;169(3):107834. doi:10.1016/j.compbiomed.2023.107834.
5. Yadav K, Alharbi Y, Alreshidi EJ, Alreshidi A, Jain AK, Jain A, et al. A comprehensive image processing framework for early diagnosis of diabetic retinopathy. Comput Mater Contin. 2024;81(2):2665–83. doi:10.32604/cmc.2024.053565.
6. Rahman A, Youldash M, Alshammari G, Sebiany A, Alzayat J, Alsayed M, et al. Diabetic retinopathy detection: a hybrid intelligent approach. Comput Mater Contin. 2024;80(3):4561–76. doi:10.32604/cmc.2024.055106.
7. Guefrachi S, Echtioui A, Hamam H. Automated diabetic retinopathy screening using deep learning. Multimed Tools Appl. 2024;83(24):65249–66. doi:10.1007/s11042-024-18149-4.
8. Khan AQ, Sun G, Khalid M, Farrash M, Bilal A. Multi-deep learning approach with transfer learning for 7-stages diabetic retinopathy classification. Int J Imaging Syst Tech. 2024;34(6):e23213. doi:10.1002/ima.23213.
9. Ding W, Sun Y, Huang J, Ju H, Zhang C, Yang G, et al. RCAR-UNet: retinal vessel segmentation network algorithm *via* novel rough attention mechanism. Inf Sci. 2024;657(3):120007. doi:10.1016/j.ins.2023.120007.
10. Liu Y, Shen J, Yang L, Yu H, Bian G. Wave-Net: a lightweight deep network for retinal vessel segmentation from fundus images. Comput Biol Med. 2023;152(2):106341. doi:10.1016/j.compbiomed.2022.106341.
11. Khan AQ, Sun G, Li Y, Bilal A, Manan MA. Optimizing fully convolutional encoder-decoder network for segmentation of diabetic eye disease. Comput Mater Contin. 2023;77(2):2481–504. doi:10.32604/cmc.2023.043239.
12. Muthusamy D, Palani P. Deep learning model using classification for diabetic retinopathy detection: an overview. Artif Intell Rev. 2024;57(7):185. doi:10.1007/s10462-024-10806-2.
13. Latif J, Tu S, Xiao C, Bilal A, Ur Rehman S, Ahmad Z. Enhanced nature inspired-support vector machine for glaucoma detection. Comput Mater Contin. 2023;76(1):1151–72. doi:10.32604/cmc.2023.040152.
14. Li J, Gao G, Yang L, Bian G, Liu Y. DPF-Net: a dual-path progressive fusion network for retinal vessel segmentation. IEEE Trans Instrum Meas. 2023;72:2517817. doi:10.1109/TIM.2023.3277946.
15. Azar AT. A bio-inspired method for segmenting the optic disc and macula in retinal images. Int J Comput Appl Technol. 2023;72(4):262–77. doi:10.1504/IJCAT.2023.133882.
16. Parsa S, Khatibi T. Grading the severity of diabetic retinopathy using an ensemble of self-supervised pre-trained convolutional neural networks: ESSP-CNNs. Multimed Tools Appl. 2024;83(42):89837–70. doi:10.1007/s11042-024-18968-5.
17. Nawaz A, Ali T, Mustafa G, Babar M, Qureshi B. Multi-class retinal diseases detection using deep CNN with minimal memory consumption. IEEE Access. 2023;11:56170–80. doi:10.1109/ACCESS.2023.3281859.
18. Tong L, Li T, Zhang Q, Zhang Q, Zhu R, Du W, et al. LiViT-Net: a U-Net-like, lightweight Transformer network for retinal vessel segmentation. Comput Struct Biotechnol J. 2024;24(8):213–24. doi:10.1016/j.csbj.2024.03.003.
19. Ma Z, Li X. An improved supervised and attention mechanism-based U-Net algorithm for retinal vessel segmentation. Comput Biol Med. 2024;168(2):107770. doi:10.1016/j.compbiomed.2023.107770.
20. Hai Z, Zou B, Xiao X, Peng Q, Yan J, Zhang W, et al. A novel approach for intelligent diagnosis and grading of diabetic retinopathy. Comput Biol Med. 2024;172(5):108246. doi:10.1016/j.compbiomed.2024.108246.

21. Cao P, Hou Q, Song R, Wang H, Zaiane O. Collaborative learning of weakly-supervised domain adaptation for diabetic retinopathy grading on retinal images. Comput Biol Med. 2022;144(99):105341. doi:10.1016/j.compbiomed.2022.105341.
22. Rodríguez MA, AlMarzouqi H, Liatsis P. Multi-label retinal disease classification using transformers. IEEE J Biomed Health Inform. 2023;27(6):2739–50. doi:10.1109/JBHI.2022.3214086.
23. Silva PS, Zhang D, Jacoba CMP, Fickweiler W, Lewis D, Leitmeyer J, et al. Automated machine learning for predicting diabetic retinopathy progression from ultra-widefield retinal images. JAMA Ophthalmol. 2024;142(3):171. doi:10.1001/jamaophthalmol.2023.6318.
24. Bodapati JD. Adaptive ensembling of multi-modal deep spatial representations for diabetic retinopathy diagnosis. Multimed Tools Appl. 2024;83(26):68467–86. doi:10.1007/s11042-024-18356-z.
25. Bondala AK, Lella KK. Revolutionizing diabetic retinopathy detection using DB-SCA-UNet with Drop Block-Based Attention Model in deep learning for precise analysis of color retinal images. Eur Phys J Spec Top. 2024;1:1–25.
26. Bilal A, Liu X, Baig TI, Long H, Shafiq M. EdgeSVDNet: 5G-enabled detection and classification of vision-threatening diabetic retinopathy in retinal fundus images. Electronics. 2023;12(19):4094. doi:10.3390/electronics12194094.
27. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021 Oct 10–17; Montreal, QC, Canada.
28. Li X, Xia H, Lu L. ECA-CBAM: classification of diabetic retinopathy: classification of diabetic retinopathy by cross-combined attention mechanism. In: Proceedings of the 2022 6th International Conference on Innovation in Artificial Intelligence; 2022 Mar 4–6; Guangzhou, China.
29. Porwal P, Pachade S, Kamble R, Kokare M, Deshmukh G, Sahasrabuddhe V, et al. Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research. Data. 2018;3(3):25. doi:10.3390/data3030025.
30. Karthik M, Dane S. Aptos 2019 blindness detection. [cited 2025 Jan 1]. Available from: https://kaggle.com/competitions/aptos2019-blindness-detection.
31. Sedik A, Kolivand H, Albeedan M. An efficient image classification and segmentation method for crime investigation applications. Multimed Tools Appl. 2024;2024(2–3):1–25. doi:10.1007/s11042-024-19773-w.
32. Zhang L, Gang J, Liu J, Zhou H, Xiao Y, Wang J, et al. Classification of diabetic retinopathy algorithm based on a novel dual-path multi-module model. Med Biol Eng Comput. 2025;63(2):365–81. doi:10.1007/s11517-024-03194-w.
33. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA.
34. Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 1–24; New Orleans, LA, USA.
35. Tan M, Le Q. Efficientnet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. In: Proceedings of the 36th International Conference on Machine Learning; 2019 Jun 9–15; Long Beach, CA, USA.
36. He A, Li T, Li N, Wang K, Fu H. CABNet: category attention block for imbalanced diabetic retinopathy grading. IEEE Trans Med Imaging. 2020;40(1):143–53. doi:10.1109/TMI.2020.3023463.
37. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. 2014.
38. Shaik NS, Cherukuri TK. Hinge attention network: a joint model for diabetic retinopathy severity grading. Appl Intell. 2022;52(13):15105–21. doi:10.1007/s10489-021-03043-5.
39. Yue G, Li Y, Zhou T, Zhou X, Liu Y, Wang T. Attention-driven cascaded network for diabetic retinopathy grading from fundus images. Biomed Signal Process Contr. 2023;80(1):104370. doi:10.1016/j.bspc.2022.104370.

40. Lei Y, Lin S, Li Z, Zhang Y, Lai T. GNN-fused CapsNet with multi-head prediction for diabetic retinopathy grading. Eng Appl Artif Intell. 2024;133(4):107994. doi:10.1016/j.engappai.2024.107994.

41. Musluh SK, Okran AM, Abdulwahab S, Puig D, Rashwan HA. Advanced diabetic retinopathy classification: integrating pathological indicators segmentation and morphological feature analysis. In: International Workshop on Ophthalmic Medical Image Analysis; 2024; Berlin/Heidelberg, Germany: Springer.