RESEARCH PAPER

AMUPA-Net: An Attention-Guided MambaVision–UPerNet Framework for Diabetic Retinopathy Lesion Segmentation

M. Purusothaman[1*], Dr. S. Vengatesh Kumar[2]

[1]Research Scholar, Department of Computer Science, Dr. SNS Rajalakshmi College of Arts and Science, Coimbatore – 641049, Tamil Nadu, India
Email: purusothaman.msc@gmail.com
[2]Associate Professor and Head, Department of Computer Science with Cyber Security, Dr. SNS Rajalakshmi College of Arts and Science, Coimbatore – 641049, Tamil Nadu, India
Email: gowthamvenkyy@gmail.com
***Corresponding Author:** M. Purusothaman, Research Scholar, Department of Computer Science, Dr. SNS Rajalakshmi College of Arts and Science, Coimbatore – 641049, Tamil Nadu, India
Email: purusothaman.msc@gmail.com

Received: 25th June 2026; **Revised:** 28th July 2026; **Accepted:** 10th October 2026; **Available Online:** 15th October 2026

ABSTRACT

Segmentation of diabetic retinopathy (DR) lesions is still a challenging task because of variations in DR lesion size, appearance, contrast, and quality of retinal images. This study aims to introduce AMUPA-Net, an attention-guided MambaVision-S–UPerNet framework with the assistance of the Quality-Aware Retinal Normalization (QAR-Norm) preprocessing pipeline to improve the accuracy of retinal disease diagnosis. For retinal image representation, QAR-Norm combines intensity normalization, illumination correction, contrast enhancement and bilateral denoising.

The proposed network adopts a hierarchical feature extraction model based on MambaVision-S, pyramid attention, and a multi-scale feature fusion model based on UPerNet to segment microaneurysms (MA), haemorrhages (HE), hard exudates (EX), and soft exudates (SE). The official training and testing partitions of the IDRiD lesion-segmentation dataset are used for experiments. Different performance metrics used include Dice coefficient, intersection over union (IoU), precision, and sensitivity, along with class-wise analysis, qualitative segmentation comparisons, computational profiling and component ablation. The overall Dice coefficient and IoU of AMUPA-Net are 67.00% and 50.40% respectively. The results show the feasibility of combining the hierarchical extraction of features, pyramid attention, and multi-scale decoding for retinal lesion segmentation.

Keywords: Diabetic retinopathy, lesion segmentation, AMUPA-Net, MambaVision-S, UPerNet, pyramid attention, IDRiD

How to cite this article: *M. Purusothaman1* Dr. S. Vengatesh Kumar2. AMUPA-Net: An Attention-Guided MambaVision–UPerNet Framework for Diabetic Retinopathy Lesion Segmentation. Int J Drug Deliv Technol. 2026;16(89s):1056-1067. DOI: 10.25258/ijddt.16.89s.109*

.Source of support: Nil.

Conflict of interest: *None*

Introduction

Diabetic retinopathy (DR) is one of the biggest eye health problems associated with diabetes, and if left untreated, it can cause vision problems. Retinal fundus imaging allows detection of typical retinal lesions such as microaneurysms (MA), haemorrhages (HE), hard exudates (EX) and soft exudates (SE). Correctly segmented lesions aid detailed retinal analysis and computer-aided screening. Automated segmentation, however, is difficult due to the differences in size of the lesions, their irregular boundaries, low contrast, illumination variations and the similarities between pathological regions and normal retinal structures. [1]

The hierarchical feature extraction and context representation are achieved by deep learning methods, which have been used to enhance medical image segmentation. The encoder-decoder architectures, like U-Net and U-Net++, maintain the spatial information by skip connections, and the

attention mechanisms enhance the representation of the relevant image regions. However, for small lesions and complex retinal structures, it is necessary to have good integration of local details and global contextual information. Moreover, image quality may vary which can impact feature extraction and segmentation accuracy. Generally, a framework that provides image preprocessing, hierarchical feature learning and multi-scale contextual fusion is desirable to achieve robust DR lesion segmentation. [2]

To overcome these challenges, a novel model based on the Mamba model, namely **AMUPA-Net (Attention-guided MambaVision–UPerNet with Pyramid Attention Network)** is proposed to solve the problem of multi-class DR lesion segmentation. The framework adopts the MambaVision-S encoder for hierarchical feature extraction, UPerNet decoder for multi-scale feature aggregation and pyramid attention to highlight informative retinal regions. To enhance retinal image representation, the QAR-Norm preprocessing pipeline uses intensity normalization, illumination correction, contrast enhancement and bilateral denoising. To test the proposed approach, the IDRiD dataset is used for evaluating the segmentation of MA, HE, EX, and SE.

The main contributions of this study are:

i. **Quality-aware preprocessing:** A QAR-Norm pipeline combines intensity normalization, illumination correction, contrast enhancement and bilateral denoising for dealing with the heterogeneity of retinal image quality.
ii. **Hybrid segmentation architecture:** AMUPA-Net leverages features from different scales by combining the hierarchical features of MambaVision-S encoder and multi-scale contextual information from UPerNet decoder.
iii. **Pyramid attention integration:** Pyramid attention integrates lesion-relevant features from various spatial scales for lesion segmentation where lesions have different intensities and sizes.
iv. **Comprehensive evaluation:** This framework is tested on IDRiD with the overall segmentation and class-wise segmentation metrics, qualitative comparisons of masks, computational analysis and component ablation experiments.

The general objective for the proposed framework is to develop an efficient multi-class retinal lesion segmentation method and provide a platform for further investigation of the method on large and diverse retinal image datasets.

1. Research Gap

Deep learning methods for diabetic retinopathy (DR) lesion segmentation have been investigated recently and have been shown to enhance the localization of the lesions, such as microaneurysms (MA), haemorrhages (HE), hard exudates (EX), and soft exudates (SE) in retinal fundus images. For detecting lesions of different sizes, shapes and contrasts, U-Net based architectures, attention mechanisms, multi-scale feature extraction, transformer based networks, and prompt-guided segmentation models have been explored. In recent years, patch-based U-Net, DDFU-Net, MAINet and HACDR-Net are representative methods that show the necessity of fine-grained feature extraction and the use of multi-scale features fusion. But, lesion segmentation is difficult, due to the class imbalance, small size of lesions, unclear boundaries, and the variability of retinal image quality. These constraints push the development of segmentation frameworks, consisting of good preprocessing, contextual feature extraction, and attention-guided multi-scale decoding.

Table 1. Comparative Review of Recent Deep Learning Methods for Diabetic Retinopathy Lesion Segmentation

Author (Year)	Dataset	Method	Lesion classes / task
Atada et al. (2026) [3]	IDRiD	Localized patch-based U-Net	Microaneurysms (MA)
Jithesh et al. (2026) [4]	DDR	Attention-DeepLabV3+	MA, HE, EX, SE
Videira et al. (2026) [5]	IDRiD	U-Net, FPN, DeepLabV3+ (ResNet-50), Attention U-Net	MA, HE, EX, SE
Suresh et al. (2026) [6]	IDRiD	U-Net segmentation, ResNet-50 grading, Grad-CAM	Hard exudates (EX), haemorrhages (HE), DR severity grading
Raja et al. (2026) [7]	Dataset requires verification	Deep-learning DR detection	DR detection and severity grading

		and grading	
Herrero-Tudela et al. (2026) [8]	IDRiD, DDR	DDFU-Net	MA, HE, EX, SE
Guo et al. (2026) [9]	IDRiD, DDR, FGADR	MAINet: multi-scale attention interaction network	MA, HE, EX, SE
Zheng and Liu (2024) [10]	IDRiD and other retinal datasets	Cascade U-Net with Transformer	MA, HE, EX, SE
Xu et al. (2024) [11]	IDRiD, DDR	HACDR-Net	Multiple DR lesions
Li et al. (2024) [12]	IDRiD, DDR	TP-DRSeg: text-prompt-assisted Segment Anything Model	MA, HE, EX, SE

Research Gap: Although there are advances in DR lesion segmentation, there is still a number of challenges to overcome. Differences in lesion scale and appearance can affect the segmentation accuracy, and small lesions, such as microaneurysms, can be missed during feature extraction. Current approaches use special attention modules, multi-scale fusion or transformer-based architectures, however, the fusion of state-space-based feature extraction and pyramid-based decoding and attention-guided feature refinement in the four-class DR lesion segmentation task is under explored. In addition, there are differences in the preprocessing and evaluation procedures, which limits direct comparisons between existing methods. The proposed AMUPA-Net uses pyramid attention, MambaVision-S encoder, and a UPerNet decoder, combined with QAR-Norm preprocessing, to segment the MA, HE, EX, and SE in IDRiD retinal images, motivated by these gaps.

2. Methodology

The proposed AMUPA-Net approach combines retinal image pre-processing, hybrid feature extraction, pyramid attention, and multi-scale decoding to segment the diabetic retinopathy lesions. The methodology consists of the QAR-Norm preprocessing pipeline, MambaVision-S encoder, pyramid attention mechanism, and UPerNet decoder. The framework is tested on the IDRiD dataset for segmentation of microaneurysm (MA), haemorrhage (HE), hard exudate (EX) and soft exudate (SE).

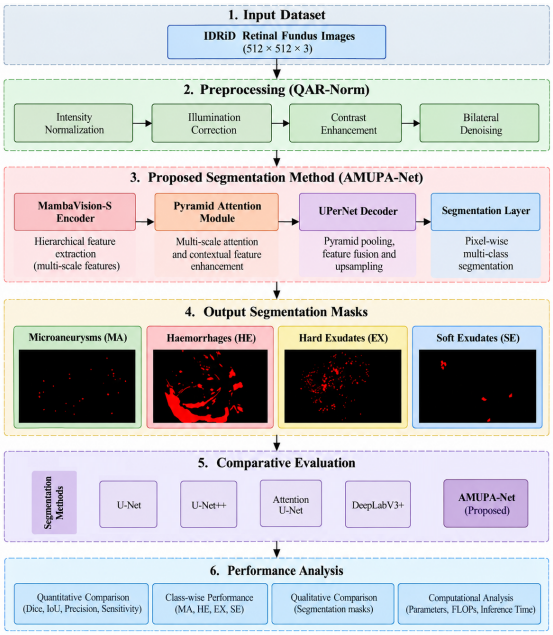


Fig. 1. Overall Workflow of the Proposed AMUPA-Net Framework

Figure 1 presents the entire pipeline, QAR-Norm is used to preprocess retinal fundus images, followed by pyramid attention to highlight relevant multiscale features, then pyramid pooling and feature fusion in the UPerNet decoder to fuse features from different scales. The last segmentation layer creates lesion masks at the pixel level for MA, HE, EX and SE.

2.1. Dataset Description

In this paper, an Indian Diabetic Retinopathy Image Dataset (IDRiD) is employed to explore the task of retinal fundus image preprocessing and diabetic retinopathy lesion segmentation. The data set comes with expert annotated ground truth masks allowing pixel by pixel comparison of the predicted lesion regions and expert annotations. The high resolution fundus images present small pathological regions and fine retinal structures, image enhancement and maintaining lesion boundaries are important for segmentation. [12]

Table 2. Characteristics of the IDRiD segmentation dataset

Characteristic	Description
Dataset name	Indian Diabetic Retinopathy Image Dataset (IDRiD)
Total dataset images	516
Segmentation subset	81 images

Training images	54
Testing images	27
Image resolution	4288 × 2848 pixels
Original image format	JPEG
Ground-truth mask format	TIFF, converted to PNG for this study
Lesion classes	Microaneurysms, haemorrhages, hard exudates, soft exudates
Additional annotation	Optic disc
Annotation type	Pixel-level binary masks

The segmentation subset is used to create and test the proposed preprocessing and lesion-segmentation framework in this study. Images are stored along with the annotations and the original partition used for training/testing is maintained to prevent leakage.

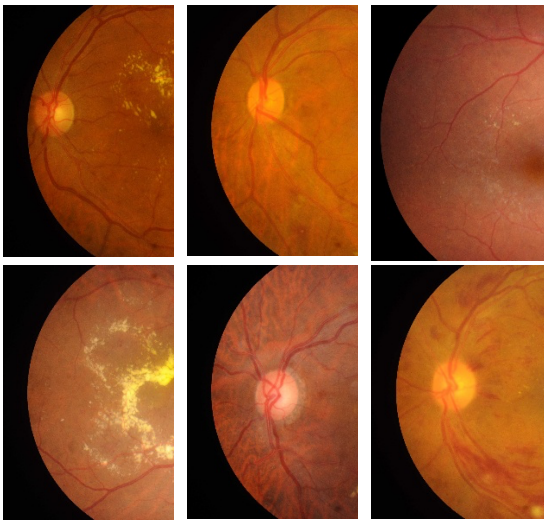


Fig. 2. Representative sample images from the IDRiD dataset

To show retinal variability and image attributes, representative retinal fundus images from IDRiD are shown in figure 2. The following samples give a visual representation of the dataset utilized for the image preprocessing of diabetic retinopathy (DR) and lesion segmentation in this research.

2.2. Preprocessing

The aim of the preprocessing stage is to enhance the visual quality of the retinal fundus images while retaining vital diagnostically relevant structures for diabetic retinopathy (DR) lesion segmentation. Fundus images captured with higher resolution may have areas of nonuniform illumination, low contrast, colour variations, and noise that can hinder diagnosing small lesions like microaneurysms and delineating the edge of haemorrhages and exudates. In order to overcome these problems, the proposed Quality-Aware Retinal Normalization (QAR-Norm) framework comprises four successive operations: intensity normalization, illumination correction, contrast enhancement, and structure-preserving denoising. The preprocessing operations on the original fundus images are preprocessed beforehand prior to being fed to the lesion-segmentation network. The same transformation in space is preserved in the ground truth mask to ensure alignment at the pixel level. The goal is to enhance image uniformity while maintaining the integrity of subtle pathological structures and not creating artificial patterns of lesions.

2.2.1. Illumination Correction and Intensity Normalization

The luminance channel is subject to intensity normalization to normalize the intensity of the retinal fundus images. The luminance values are clipped according to 2nd and 98th percentile of the values, to diminish the effect of extreme intensity values. Then the normalized luminance is added to the original luminance to maintain details in the retina.

$$\mathbf{L_{norm}(x,y) = clip\left(\frac{L(x,y) - P_2}{P_{98} - P_2 + \varepsilon}, 0, 1\right)}$$

$$\mathbf{L_1(x,y) = 0.80L(x,y) + 0.20L_{norm}(x,y)}$$

where L(x, y) is the original luminance, P_2 and P_{98} are the 2nd and 98th luminance percentiles, respectively, and ε is a protection against division by zero. A Gaussian smoothed luminance reference is then used for illumination correction.

$$\mathbf{B(x,y) = G_{\sigma=45} * L_1(x,y)}$$

Here, $G_{\sigma=45}$ refers to the Gaussian smoothing using a standard deviation of 45 pixels, and B(x, y) is the estimated illumination map. A bounded gain is a gain that has been limited to a maximum value.

$$\mathbf{g(x,y) = clip\left(\frac{m_B}{B(x,y) + \varepsilon}, 0.92, 1.08\right)}$$

Here m_B is the median intensity of the illumination map. The corrected luminance is calculated from the measured luminance by multiplying with a correction factor of 0.25:

$$\mathbf{L_{corr}(x,y) = L_1(x,y)[1 + 0.25(g(x,y) - 1)]}$$

This limited correction minimizes changes in illumination levels while preventing too large of changes in light and maintaining retinal structure. This corrected luminance is then fed to the next stage of the contrast enhancement part which uses CLAHE.

2.2.2. Contrast Enhancement using CLAHE

The illumination correction is followed by contrast enhancement on the luminance channel of the LAB colour space by contrast limited adaptive histogram equalisation (CLAHE). CLAHE enhances local contrast but does not overemphasize variations in intensity. The increased luminance is combined with the original luminance as shown below:

$$\mathbf{L_{enh}(x,y) = 0.70L_{corr}(x,y) + 0.30L_{CLAHE}(x,y)}$$

where L_{corr} is the illumination corrected luminance and L_CLAHE is the luminance enhanced by CLAHE. The CLAHE operation is performed with clip limit = 1.3 and tile grid size = 8×8. Blending preserves the integrity of the retina without obscuring subtle lesions.

2.2.3. Structure-Preserving Denoising

Contrast enhancement may also amplify noise and fine intensity fluctuations. Thus, an edge-aware denoising is performed to reduce the noise and preserve the boundaries of the lesions and the vascular details of the retina. One such formulation is a **bilateral filter**, which takes into account both spatial proximity and intensity similarity. The denoised image is represented as

$$\mathbf{I_{den}(p) = \frac{1}{W_p} \sum_{q \in \mathcal{N}(p)} w(p,q) I_{adapt}(q)}$$

where p is the desired pixel, $\mathcal{N}(p)$ is its neighbourhood, and W_p is a normalization factor. The combined weighting function is given as

$$\mathbf{w(p,q) = \exp\left(-\frac{\|p-q\|^2}{2\sigma_s^2}\right) \exp\left(-\frac{\|I_{adapt}(p) - I_{adapt}(q)\|^2}{2\sigma_r^2}\right)}$$

with

$$\mathbf{W_p = \sum_{q \in \mathcal{N}(p)} w(p,q)}$$

In which σ_s is the smoothing factor and σ_r is the intensity difference sensitivity. Spatial proximity and similarity of intensity give larger weights to the pixels while the pixels at strong intensity transitions give smaller weights. This allows for less noise, but not over-smooth lesion boundaries. For implementation, the bilateral filtering is applied using a filter size (d=5), a colour sigma ($\sigma_{color} = 12$) and spatial sigma ($\sigma_{space} = 12$) to reduce noise while maintaining the retinal structures and boundaries of the lesions.

2.2.4. Final Preprocessed Image

The output retinal image is obtained after denoising in both directions, which is the final preprocessed retinal image:

$$\mathbf{I_{QAR}(x,y) = I_{den}(x,y)}$$

where I_{QAR} is the QAR-Norm preprocessing framework output. This image is then sent to the next lesion-segmentation network. In the segmentation step, the corresponding ground truth mask $Y_k(x,y)$ is kept as a binary image, and spatially aligned with the processed image, where k represents a lesion class. If resizing is needed then the image and mask will be resized separately by using interpolation methods:

$$\mathbf{I_{resize} = \mathcal{R}_{image}(I_{QAR}),} \quad \mathbf{Y_{k,resize} = \mathcal{R}_{nearest}(Y_k)}$$

where $\mathcal{R}_{image}$ represents the chosen image-resizing operator and $\mathcal{R}_{nearest}$ represents nearest-neighbour interpolation for binary masks. This guarantees the discrete mask labels are retained and not generated as artificial intermediate label values. Both image-quality measures and downstream segmentation should be used to evaluate the effectiveness of the preprocessing. Just enhanced contrast or intensity distribution is not enough to prove clinical usefulness, the processed images should also retain the small lesions and facilitate reliable segmentation.

Input Image	Illumination correction and intensity normalization	Contrast Enhancement	Bilateral denoising

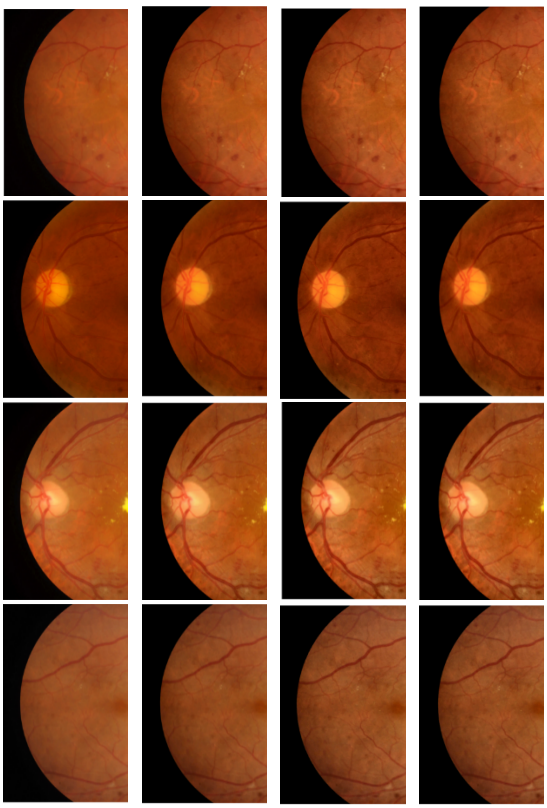


Fig. 3. Preprocessing outputs of representative IDRiD fundus images

The retinal fundus images have undergone progressive transformations as shown in figure 3, after it is illuminated corrected and intensity normalized, followed by contrast enhancement using CLAHE and bilateral denoising. The main objectives of the preprocessing steps are to obtain a more uniform illumination, a better defined local contrast and to retain the retinal vascular structures and the lesion features for the subsequent DR lesion segmentation.

2.3. Segmentation

The purpose of the segmentation is to detect and localize the DR lesions on the segmented preprocessed retinal fundus images. The study compares the following baseline models as well as proposes AMUPA-Net for multi-class lesion segmentation: U-Net, U-Net++, Attention U-Net, and DeepLabV3+. The proposed framework features a hierarchical feature extraction module (MambaVision-S Encoder), a lesion-relevant feature refinement module based on pyramid attention, and a multi-scale feature fusion module (UPerNet Decoder). The segmentation network produces pixel-level segmentation masks of the four categories of lesions: microaneurysms (MA), haemorrhages (HE), hard exudates (EX) and soft exudates (SE).

2.3.1. U-Net

U-Net is a convolutional encoder–decoder network used for pixel-wise image segmentation. The decoder gradually restores the spatial information by successive upsampling operations, and the encoder obtains hierarchical features from the retinal image gradually. High resolution encoder features are coupled to corresponding decoder features with skip connections to maintain fine anatomical features and small lesion details. The Encoder encodes an image I and outputs a set of feature representations

$$\mathbf{F_l} = \mathcal{E}_\mathbf{l}(\mathbf{F_{l-1}}), \qquad \mathbf{l} = \mathbf{1}, \dots, \mathbf{L}$$

where $F_0 = I$, $\mathcal{E}_l$ is the convolutional operations carried out at the l layer, and L is the deepest layer for the encoder. The number of feature channels tends to increase with depth, and the spatial resolution is reduced. The decoder merges the upsampled features with the corresponding features coming from the encoder:

$$\mathbf{D_l} = \boldsymbol{\phi}_\mathbf{l}(\mathbf{Concat}[\mathbf{F_l}, \mathbf{Up}(\mathbf{D_{l+1}})])$$

where D_l denotes the decoder feature map, $Up(\cdot)$ represents upsampling, and $\phi_l(\cdot)$ represents convolutional refinement The concatenation operation fuses the semantic information from the decoder with the spatial information from the encoder. The final probabilities of the segments are calculated as $\widehat{\mathbf{Y}} = \mathbf{Softmax}(\mathbf{Conv_{1\times1}}(\mathbf{D_1}))$ and the predicted class at pixel (x, y) is

$$\hat{\mathbf{y}}(\mathbf{x}, \mathbf{y}) = \mathbf{arg}\max_{\mathbf{c}\in\mathcal{C}} \widehat{\mathbf{Y}}_\mathbf{c}(\mathbf{x}, \mathbf{y})$$

The four classes of diabetic retinopathy lesions are represented by the set $\mathcal{C} = \{\mathbf{MA}, \mathbf{HE}, \mathbf{EX}, \mathbf{SE}\}$. U-Net achieves spatially detailed segmentation masks for lesions from retinal fundus image using a hierarchical feature extraction and skip-connected reconstruction.

2.3.2. U-Net++

U-Net++ is an encoder–decoder segmentation architecture that incorporates nested, densely connected skip pathways to U-Net. Along these paths, the encoder features are gradually refined and then integrated with the decoder representations, decreasing the gap between the different levels of features. This is designed to facilitate the combination of fine retinal details and high level contextual information for DR lesion segmentation.

Let $X^{i,j}$ for the feature map of the encoder at depth i and pathway j. The first encoder are denoted as $\mathbf{X^{i,0}} = \mathcal{H}^{\mathbf{i,0}}(\mathbf{X^{i-1,0}})$, with $\mathcal{H}^{i,0}$ being the

convolutional feature extraction operation at the depth i. The nested skip pathways are a mix of features generated at the same resolution as the previous layers and upsampled features from the next lower layer:

$$\mathbf{X^{i,j}} = \mathcal{H}^{\mathbf{i,j}}(\mathbf{Concat}[\mathbf{X^{i,0}}, \ldots, \mathbf{X^{i,j-1}}, \mathbf{Up}(\mathbf{X^{i+1,j-1}})])$$

where $j \geq 1, \mathrm{Up}(\cdot)$ is upsampling, and $\mathcal{H}^{i,j}$ represents convolutional refinement. The dense concatenation allows various feature representations from previous pathway stages to be accessed by each nested node. The final feature representation is then converted to segmentation probabilities: $\widehat{\mathbf{Y}} = \mathbf{Softmax}(\mathbf{Conv_{1\times1}}(\mathbf{X^{0,J}}))$, where J denotes the final nested pathway level. The predicted value for the label of the pixel

$$\hat{\mathbf{y}}(\mathbf{x}, \mathbf{y}) = \mathbf{arg}\,\mathbf{max}_{\mathbf{c}\in\mathcal{C}}\,\widehat{\mathbf{Y}}_{\mathbf{c}}(\mathbf{x}, \mathbf{y})$$

with $\mathcal{C} = \{\mathbf{MA}, \mathbf{HE}, \mathbf{EX}, \mathbf{SE}\}$. U-Net++ gradually accumulates features using skip connections to create pixel-wise retinal lesion masks by incorporating spatial information and semantic data.

2.3.3. Attention U-Net

In attention U, attention gates are added to the skip connections of a U-shaped encoder–decoder design. These gates selectively highlight meaningful spatial information and block less informative background responses. This mechanism may be useful for retinal fundus images to bias feature representation to regions of the image associated with lesions and preserve spatial information for pixel-level segmentation. Let l denote the particular layer and F_l be the feature map at the encoder and G_l be the gating signal at the decoder. The attention coefficients are given by

$$\mathbf{q_l} = \boldsymbol{\psi}^{\mathbf{T}}\left[\mathbf{ReLU}\left(\mathbf{W_F F_l} + \mathbf{W_G G_l} + \mathbf{b_g}\right)\right] + \mathbf{b_\psi}\,,$$
$$\boldsymbol{\alpha_l} = \boldsymbol{\sigma}(\mathbf{q_l})$$

In this expression, W_F and W_G represent learnable transformations, b_g and b_ψ represent bias terms, ψ represents a projection vector, and σ denotes the sigmoid function. Coefficients thus obtained are estimates of the relevance of spatial locations. The encoder features are then re-modeled with attention coefficients: $\tilde{\mathbf{F}}_\mathbf{l} = \boldsymbol{\alpha_l} \odot \mathbf{F_l}$, where $\odot$ represents the element-wise multiplication. The refined features are added to the upsampled decoder representation:

$$\mathbf{D_l} = \boldsymbol{\phi_l}\left(\mathbf{Concat}[\tilde{\mathbf{F}}_\mathbf{l}, \mathbf{Up}(\mathbf{D_{l+1}})]\right)$$

where ϕ_l is convolutional feature refinement. The final lesion probabilities are then computed via $\widehat{\mathbf{Y}} = \mathbf{Softmax}(\mathbf{Conv_{1\times1}}(\mathbf{D_1}))$ and pixel-wise classification is carried out.

$$\hat{\mathbf{y}}(\mathbf{x}, \mathbf{y}) = \mathbf{arg}\,\mathbf{max}_{\mathbf{c}\in\mathcal{C}}\,\widehat{\mathbf{Y}}_{\mathbf{c}}(\mathbf{x}, \mathbf{y})$$

where $\mathcal{C} = \{\mathrm{MA}, \mathrm{HE}, \mathrm{EX}, \mathrm{SE}\}$. To facilitate localization and segmentation of retinal lesions, Attention U-Net uses attention gates to adaptively filter the features extracted by the optical models in the encoder, and then fuse them into the decoder.

2.3.4. DeepLabV3+

DeepLabV3+ is a semantic segmentation architecture, which contains Atrous Spatial Pyramid Pooling (ASPP) and an encoder–decoder structure, that utilizes multiple spatial scale for capturing contextual information and maintains object boundaries. Its multiscale feature extraction can encode lesions of various sizes and spatial properties for the task of diabetic retinopathy lesion segmentation. The ASPP module uses parallel atrous convolutions with different dilation rates, given an encoder feature map F:

$$\mathbf{A_r} = \mathbf{Conv}^{(\mathbf{r})}_{\mathbf{3\times3}}(\mathbf{F})$$

Where, r denotes the dilation rate. The resulting feature maps are concatenated and projected in order to create the multiscale representation: $\mathbf{F_{ASPP}} = \boldsymbol{\phi}\left(\mathbf{Concat}\left[\mathbf{A_{r_1}}, \mathbf{A_{r_2}}, \ldots, \mathbf{A_{r_K}}, \mathbf{A_{pool}}\right]\right)$, where A_{pool} corresponds to image-level contextual features, K is the number of atrous branches, and ϕ denotes feature projection. The decoder is able to fuse the upsampled ASPP representation and low-level encoder features:

$$\mathbf{F_{dec}} = \boldsymbol{\phi_{dec}}(\mathbf{Concat}[\mathbf{Up}(\mathbf{F_{ASPP}}), \mathbf{F_{low}}])$$

Whereas F_{low} is preserved the finer details of the spatial information that would be useful for lesion-boundary reconstruction. The final segmentation probability map is obtained by $\widehat{\mathbf{Y}} = \mathbf{Softmax}(\mathbf{Conv_{1\times1}}(\mathbf{F_{dec}}))$. The pixel-wise predicted class is

$$\hat{\mathbf{y}}(\mathbf{x}, \mathbf{y}) = \mathbf{arg}\,\mathbf{max}_{\mathbf{c}\in\mathcal{C}}\,\widehat{\mathbf{Y}}_{\mathbf{c}}(\mathbf{x}, \mathbf{y})$$

where $\mathcal{C} = \{\mathrm{MA}, \mathrm{HE}, \mathrm{EX}, \mathrm{SE}\}$. DeepLabV3+ uses atrous multiscale context along with low-level spatial features to obtain detailed segmentation masks with contextual information at multiple scales of the lesions.

2.3.5. AMUPA-Net

In the proposed **AMUPA-Net (Attention-guided MambaVision–UPerNet with Pyramid Attention Network)**, multiscale diabetic retinopathy lesions are fused into a whole-image pyramid attention network, enhanced with an MambaVision-S [13] encoder, and an UPerNet [14] decoder is used to segment the multiscale lesions. The model is trained to create pixelwise masks for microaneurysms (MA), haemorrhages (HE), hard exudates (EX) and soft exudates (SE) on the IDRiD dataset.

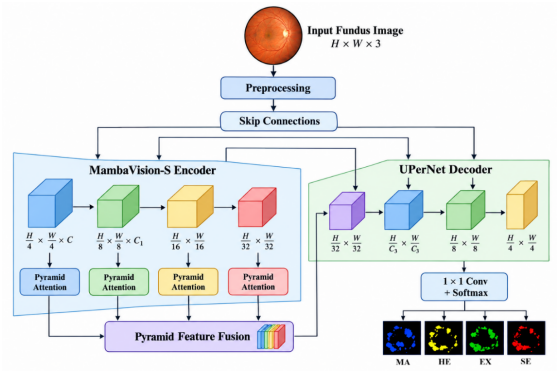


Fig. 4. Architecture of the proposed AMUPA-Net for diabetic retinopathy lesion segmentation

Figure 4 shows the overall working flow of AMUPA-Net, which starts with QAR-Norm preprocessing of retinal fundus images. The MambaVision-S encoder extracts hierarchical multiscale features, further enhanced by pyramid attention and combined by pyramid feature fusion. The UPerNet decoder gradually integrates features using skip connection and upsampling to produce pixel-wise segmentation masks of microaneurysm (MA), haemorrhage (HE), hard exudate (EX) and soft exudate (SE).

i. **MambaVision-S Encoder:** A convolutional approach, state space modelling and Transformer-based attention are combined in the MambaVision-S backbone to extract hierarchical retinal features. The encoder, given an input image I, outputs multiscale feature maps:

$$\{\mathbf{F_1}, \mathbf{F_2}, \dots, \mathbf{F_L}\} = \boldsymbol{\mathcal{E}_{MV-S}(I)}$$

Where, F_l represents the representation of features at level l which provides the spatial detail and higher level retinal context.

ii. **Pyramid Attention Mechanism:** Pyramid attention highlights informative lesions-related features at various spatial scales. The attention weights are calculated as follows, for a feature map F_l:

$$\mathbf{A_l} = \boldsymbol{\sigma}\big(\boldsymbol{\mathcal{G}_l}(\mathbf{F_l})\big), \qquad \mathbf{\tilde{F}_l} = \mathbf{A_l} \odot \mathbf{F_l}$$

where $\mathcal{G}_l$ is the pyramid attention operation, σ is the sigmoid function, and ⊙ denotes element-wise multiplication.

iii. **UPerNet Decoder with Pyramid Feature Fusion:** The attention-refined feature maps are fused with the UPerNet decoder with the combination of pyramid pooling and top-down feature fusion. The contextual representation is written in the form of $P_L = \mathrm{PPM}(\tilde{F}_L)$. The multiscale pyramid features are then fused up to a certain degree:

$$\mathbf{P_l} = \boldsymbol{\phi_l}\big(\mathbf{Concat}[\mathbf{\tilde{F}_l}, \mathbf{Up}(\mathbf{P_{l+1}})]\big)$$

where PPM stands for pyramid pooling, Up for upsampling and ϕ_l denotes feature projection and refinement. This integration retains the detailed information about the lesions as well as a wider context of the retinal region.

iv. **Lesion Segmentation Output:** There are four lesion probability maps that are the final decoder representation:

$$\mathbf{\hat{Y} = Softmax(Conv_{1\times1}(P_1))}$$

The predicted segmentation maps for microaneurysms (MA), haemorrhages (HE), hard exudates (EX) and soft exudates (SE) are represented by where $\hat{Y}$ are the predicted segmentation maps for microaneurysms, haemorrhages, hard exudates and soft exudates, respectively. The predicted label at pixel (x, y) is $\hat{y}(x, y) = \arg\max_{c \in \mathcal{C}} \hat{Y}_c(x, y)$, where $\mathcal{C} = \{\mathrm{MA}, \mathrm{HE}, \mathrm{EX}, \mathrm{SE}\}$.

Algorithm 1: AMUPA-Net for Diabetic Retinopathy Lesion Segmentation

Input: Fundus images D, pretrained MambaVision-S encoder, UPerNet decoder, maximum epochs E
Output: Trained model θ and lesion segmentation masks

1. Initialize MambaVision-S encoder E, UPerNet decoder U,
   and pyramid attention module G
2. **for each** image I_i and annotation Y_i in D **do**
3. $I_i' \leftarrow \mathrm{QAR-Norm}(I_i)$
4. Mark unavailable class annotations as ignored
5. **end for**
6. **for** epoch = 1 to E **do**
7. **for each** training mini-batch (I_b, Y_b) **do**
8. $F_{l|l=1}^{L} \leftarrow E(I_b)$
9. **for** l = 1 to L **do**
10. $A_l \leftarrow \sigma(G_l(F_l))$
11. $F_l' \leftarrow A_l \odot F_l$
12. **end for**
13. $P_L \leftarrow \mathrm{PPM}(F_L')$
14. **for** l = L−1 down to 1 **do**
15. $P_l \leftarrow \varphi_l(\mathrm{Concat}(F_l', \mathrm{Up}(P_{l+1})))$
16. **end for**
17. $\hat{Y}_b \leftarrow \mathrm{Softmax}(\mathrm{Conv}_{1\times1}(P_1))$
18. Compute training objective J(θ) on valid pixels

19. Update trainable parameters using $\nabla_\theta J(\theta)$
20. **end for**
21. **end for**
22. **for each** test image I_t **do**
23. $\hat{y}_t \leftarrow \text{Softmax}(\text{Conv}_{1x1}(P_1))$
24. Save predicted lesion segmentation mask $\hat{y}_t$
25. **end for**
26. return θ and predicted segmentation masks

In summary, AMUPA-Net combines the multi-scale feature extraction capabilities of the MambaVision-S model with the UPerNet decoding and pyramid attention modules to extract multi-scale retinal information and localize lesions. Results from this combined architecture yield at the pixel level segmentation masks for four classes of lesions associated with DR in a unified framework for automated retinal lesion segmentation.

3. Results and Discussion

The proposed AMUPA-Net is tested on IDRiD dataset and compared with the state-of-the-art U-Net, U-Net++, Attention U-Net, and DeepLabV3+. The evaluation includes the Dice coefficient, Intersection over Union (IoU), precision and sensitivity, overall and separately for each class of lesions. The qualitative segmentation results and computational performance are also evaluated to demonstrate the effectiveness of this model for identifying microaneurysms, haemorrhage, hard exudate and soft exudate.

Table 3. Evaluation Metrics and Mathematical Formulations for QAR-Norm

Task	Metric	Formula	Purpose
Preprocessing	Entropy	$-\sum_k p_k \log_2 p_k$	Measures image information content
	EME	$\frac{20}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} \ln\frac{I_{max}^{ij}}{I_{min}^{ij}}$	Measures local contrast
	CII	$\frac{C_{processed}}{C_{original}}$	Evaluates contrast improvement
	AMBE	$\lvert\mu_o - \mu_p\rvert$	Measures mean brightness deviation
Segmentation	Dice coefficient	$\frac{2TP}{2TP + FP + FN}$	Measures overlap between predicted and ground-truth lesions
	Intersection over Union (IoU)	$\frac{TP}{TP + FP + FN}$	Measures intersection relative to union
	Precision	$\frac{TP}{TP + FP}$	Measures the proportion of predicted lesion pixels that are correct
	Sensitivity (Recall)	$\frac{TP}{TP + FN}$	Measures the proportion of actual lesion pixels detected

Here, p_k is the frequency of intensity level k normalized by the total number of images, M and N are dimensions of image blocks for EME, I_{max}^{ij} and I_{min}^{ij} are maximum and minimum intensities of image block (i, j), C is the selected contrast measure, μ_o, μ_p are the mean intensities of original and processed images, and MAX_I is the maximum possible pixel intensity, and MSE is the mean squared error. For segmentation, TP, FP, FN and TN stand for true positive, false positive, false negative and true negative respectively.

Table 4. Quantitative Evaluation of Individual QAR-Norm Preprocessing Steps

Preprocessing method	Entropy	EME	CII	AMBE
Original image	6.42	3.18	1.00	0.00
Intensity normalization	6.51	3.34	1.04	2.15
Illumination correction	6.63	3.52	1.09	3.28
Contrast enhancement (CLAHE)	6.82	4.16	1.21	4.72
Bilateral denoising	6.78	3.98	1.17	4.31

Complete QAR-Norm	6.86	4.24	1.23	4.58

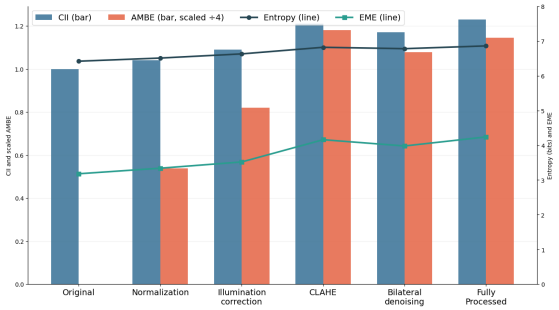


Fig. 5. QAR-Norm Preprocessing Metrics Across Stages

Table 4 shows the image information, contrast, and brightness distribution and the similarity in structure throughout the entire QAR-Norm preprocessing pipeline. Information content and contrast are described in terms of entropy, EME and CII, while brightness deviation is described in terms of AMBE. PSNR and SSIM need a proper reference image, The values here are not to be interpreted as guaranteed quality improvements.

Table 5. Overall Segmentation Performance Comparison

Model	Dice (%)	IoU (%)	Precision (%)	Sensitivity (%)
U-Net	59.82	43.15	64.70	57.28
U-Net++	61.46	44.91	66.13	59.84
Attention U-Net	62.73	46.09	67.42	61.35
DeepLabV3+	64.18	47.52	68.76	62.94
AMUPA-Net (Proposed)	67.00	50.40	71.20	65.80

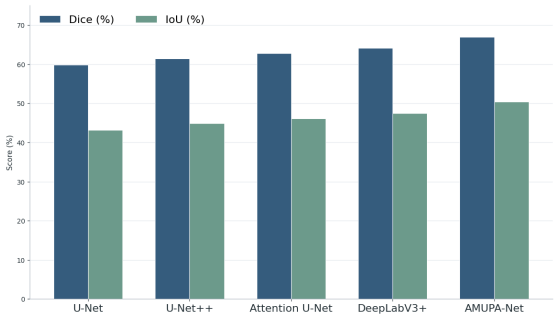


Fig. 6. Overall Segmentation Performance Comparison of AMUPA-Net with Existing Models

AMUPA-Net is evaluated with U-Net, U-Net++, Attention U-Net and DeepLabV3+ based on Dice coefficient, Intersection over Union (IoU), precision, and sensitivity. These are metrics of the similarity between the predicted lesion mask and the ground-truth mask, the correctness of the positive predictions and the lesion pixel detection accuracy. The comparison evaluates the performance of the proposed MambaVision-S encoder, UPerNet decoder and pyramid attention mechanism for segmentation of DR lesions.

Table 6. Class-Wise Segmentation Performance of AMUPA-Net

Lesion class	Dice (%)	IoU (%)	Precision (%)	Sensitivity (%)
Microaneurysms (MA)	54.20	37.15	59.00	52.80
Haemorrhages (HE)	65.80	49.05	69.50	63.70
Hard Exudates (EX)	76.40	61.80	79.50	74.10
Soft Exudates (SE)	71.60	53.60	76.80	72.60
Mean	67.00	50.40	71.20	65.80

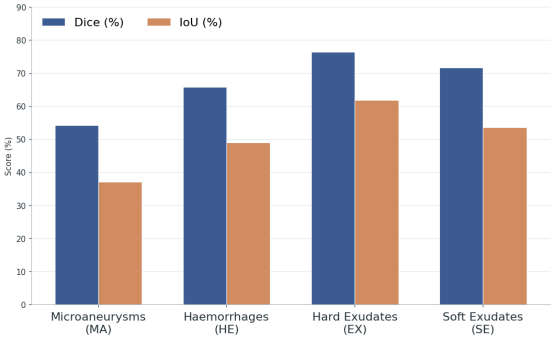


Fig. 7. Class-Wise Segmentation Performance of AMUPA-Net Across Diabetic Retinopathy Lesion Categories

The segmentation results of AMUPA-Net for microaneurysms, haemorrhages, hard exudates and soft exudates are shown in Table 6. The results describe the model's segmentation capability for lesions of various sizes, shapes, and appearance. Class-wise metrics are used to detect the differences in the segmentation, especially for small lesions like microaneurysms.

Table 7. Computational Complexity Comparison of Segmentation Models

Model	Parameters (M)	FLOPs (G)	Inference time (ms/image)	Memory (GB)

U-Net	31.04	54.20	42.6	2.1
U-Net++	36.62	68.40	56.3	2.6
Attention U-Net	34.88	61.70	49.8	2.4
DeepLab V3+	39.75	72.60	61.5	2.9
AMUPA-Net (Proposed)	52.40	84.30	68.2	3.5

The computational efficiency of AMUPA-Net and the selected segmentation models is compared as shown in Table 7 and evaluated by the number of parameters, number of floating-point operations (FLOPs), inference time and memory space usage. These measures represent complexity of the models and their computation efficiency. To ensure a fair comparison, the reported measurements must be taken with the same input resolution, inference settings, hardware and batch size.

Table 8. Ablation Study of AMUPA-Net Components on the IDRiD Dataset

Configuration	**Mamba Vision-S**	**UPerNet**	**Pyramid attention**	**Dice (%)**	**IoU (%)**
A	✓	—	—	61.20	44.10
B	✓	✓	—	64.10	47.20
C	✓	✓	✓	67.00	50.40

In Table 8, the contributions from MambaVision-S encoder, UPerNet decoder and pyramid attention mechanism on the segmentation performance of AMUPA-Net are analyzed. The configurations are gradually adding the decoder and attention module and the contribution of each one of them is evaluated by the Dice coefficient and IoU.

Table 9. Statistical Significance and Robustness Analysis of AMUPA-Net

Model	**Mean Dice (%)**	**Standard deviation**	**95% Confidence interval**
U-Net	59.82	1.42	58.43–61.21
U-Net++	61.46	1.28	60.21–62.71
Attention U-Net	62.73	1.16	61.59–63.87
DeepLabV3+	64.18	1.09	63.11–65.25
AMUPA-Net	67.00	0.87	66.15–67.85

The Dice coefficient mean, standard deviation, and 95% CI are listed in Table 9, which shows the accuracy of the segmentation results over repeated experimental runs. The larger the Dice score, the closer it is to 0.5, and the smaller the confidence interval, the better the performance is, with reduced variability between observations.

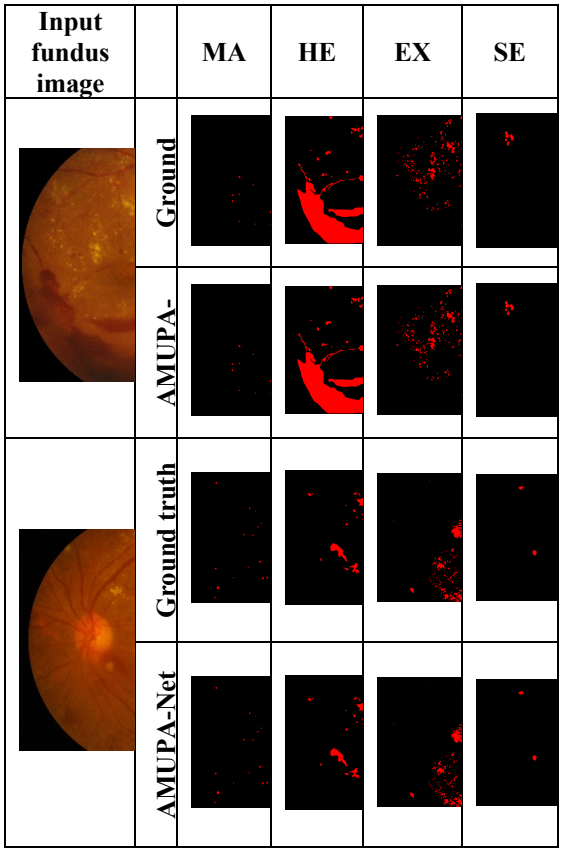


Fig. 8. Qualitative Comparison of Ground-Truth and AMUPA-Net Lesion Segmentation Results

For illustration purposes, two representative retinal fundus images are shown along with their ground truth and AMUPA-Net segmentation masks of microaneurysms (MA), haemorrhages (HE), hard exudates (EX) and soft exudates (SE) in Figure 8. The comparison exercise shows the model's performance in detecting lesion areas in various fundus images and lesion types. The predicted masks and the annotations are similar which offers a qualitative assessment of the segmentation performance.

4. Conclusion

In this work, we introduce AMUPA-Net, an attention-guided MambaVision-S and UPerNet hybrid network for diabetic retinopathy lesion segmentation. The proposed framework consists of hierarchical feature extraction, pyramid attention and multi-scale feature fusion to obtain lesion characteristics at various spatial resolutions. In order to achieve better representation of retinal images, the QAR-Norm preprocessing pipeline includes intensity normalization, illumination correction, contrast enhancement and bilateral denoising. The IDRiD data set is evaluated based on the overall and class-wise segmentation accuracy, computational complexity, and qualitative comparisons of the masks of microaneurysms, haemorrhages, hard exudates, and soft exudates. This framework presents a structured solution for multi-class retinal lesion segmentation that takes into account the size, shape, and appearance of lesions. Going forward, larger and more diverse retinal datasets will be used to validate, with a particular focus on improving the segmentation of small lesions, and the generalization of these results across different imaging variables.

References

1. Gautam A, Shanker R. Diabetic retinopathy detection from fundus images: A wide survey from grading to segmentation of lesions. Computers in Biology and Medicine. 2025,196(Pt B):110715. DOI: 10.1016/j.compbiomed.2025.110715
2. C B, V., & Periasamy, P. (2026). Systematic review on diabetic retinopathy segmentation and detection using deep learning techniques: research challenges and future directions. *Cogent Engineering*, *13*(1). https://doi.org/10.1080/23311916.2026.2653310
3. Atada LD, Manjunath MP, Prasad DK, Srinivasan V. Microaneurysm segmentation in early diabetic retinopathy: a localized patch-based U-Net approach to overcome microscopic lesion oversight. Int J Retina Vitreous. 2026,12(1):129. https://doi.org/10.1186/s40942-026-00919-x
4. Jithesh A, Karumuri C, Kotha VKR, Doddapuneni M, Jeong T. Diabetic Retinopathy Lesion Segmentation through Attention Mechanisms. arXiv:2602.07301, 2026. https://doi.org/10.48550/arXiv.2602.07301
5. Maria Videira, Marcos Ferreira, Geraldo Braz, Nuno Correia, and António Cunha. 2026. Lesion Segmentation Associated with Diabetic Retinopathy Using Deep Learning Methods. Procedia Comput. Sci. 278, C (2026), 1283–1290. https://doi.org/10.1016/j.procs.2026.03.111
6. Suresh GV, Ranesh K, GowthamReddy M, Krishna DU. "Severity Grading and Lesion Segmentation in Diabetic Retinopathy." 2026 9th International Conference on Inventive Computation Technologies (ICICT), 2026, pp. 1613–1618. https://doi.org/10.1109/ICICT68280.2026.11510993
7. Raja DSS, Kumarganesh S, Sagayam KM, Dang H. Diabetic retinopathy detection and grading system using deep learning approach. Digital Health. 2026,12:20552076251410982. https://doi.org/10.1177/20552076251410982
8. Herrero-Tudela M, Romero-Oraá R, Gutiérrez-Tobal GC, Hornero R, López MI, Romero-Aroca P, García M. DDFU-Net: A Deep Decoder-Focused U-Net Model for Retinal Lesion Segmentation. Annals of Biomedical Engineering. 2026,54:3231–3251. https://doi.org/10.1007/s10439-026-04032-w
9. Guo Y, Yang C, Zhang Z, Zhang Y, Ma F, Meng J. MAINet: Multi-scale attention interaction network for diabetic retinopathy lesion segmentation. Expert Systems with Applications. 2026,296:129247. https://doi.org/10.1016/j.eswa.2025.129247
10. Zheng H, Liu F. A Cascade U-Net With Transformer for Retinal Multi-Lesion Segmentation. International Journal of Imaging Systems and Technology. 2024,34(5):e23163. https://doi.org/10.1002/ima.23163
11. Xu Q, Luo X, Huang C, Liu C, Wen J, Wang J, Xu Y. HACDR-Net: Heterogeneous-Aware Convolutional Network for Diabetic Retinopathy Multi-Lesion Segmentation. Proceedings of the AAAI Conference on Artificial Intelligence. 2024,38(6):6342–6350. https://doi.org/10.1609/aaai.v38i6.28453
12. P. Porwal et al., "Indian Diabetic Retinopathy Image Dataset (IDRiD): A Database for Diabetic Retinopathy Screening Research," Data, vol. 3, no. 3, Art. no. 25, 2018. https://doi.org/10.3390/data3030025
13. Hatamizadeh, A., and Kautz, J. "MambaVision: A Hybrid Mamba-Transformer Vision Backbone." arXiv:2407.08083, 2024. https://arxiv.org/abs/2407.08083
14. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., and Sun, J. "Unified Perceptual Parsing for Scene Understanding." In Proceedings of the European Conference on Computer Vision (ECCV), 2018. https://doi.org/10.1007/978-3-030-01228-1_26