Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.083951

ARTICLE

Enhancing U-Net for Optic Cup and Disc Segmentation in Retinal Images Using Atrous Spatial Pyramid Pooling, Inception Modules, and Attention Gates

Anita Desiani[1,*], Indri Ramayanti[2], Sigit Priyanta[3], Bambang Suprihatin[1], Muhammad Arhami[4], Deshinta Arrova Dewi[5] and Puspa Sari[1]

[1]Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Sriwijaya, Inderalaya, Indonesia
[2]Department of Medicine, Faculty of Medicine, Universitas Muhammadiyah, Palembang, Indonesia
[3]Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Yogyakarta, Indonesia
[4]Department of Informatics Engineering, Politeknik Negeri Lhokseumawe, Aceh, Indonesia
[5]Faculty of Data Science and Information Technology, INTI International University, Nilai, Malaysia

*Corresponding Author: Anita Desiani. Email: anita_desiani@unsri.ac.id

Received: 14 April 2026; Accepted: 14 May 2026; Published: 30 June 2026

ABSTRACT: Image segmentation is essential in medical image analysis for glaucoma screening. Accurate delineation of the optic disc (OD) and optic cup (OC) in retinal fundus images is required for reliable clinical assessment. Manual segmentation is time-consuming and suffers from interobserver variability, which leads to inconsistent results. To address these limitations, this study proposes ASPP Inception Attention U-Net (ASPPIAU-Ne), an enhanced encoder-decoder architecture that integrates Atrous Spatial Pyramid Pooling (ASPP), Inception modules, and attention gates for feature selection in skip connections. The ASPP module is applied after the encoder to capture multi-scale contextual information and improve the representation of global and local structures. Attention gates suppress irrelevant features in skip connections and enhance important anatomical regions. In the decoder, Inception modules improve feature reconstruction and reduce upsampling artifacts. The proposed model is evaluated on DRISHTI-GS and REFUGE for optic disc, optic cup, and background segmentation. On the REFUGE dataset, the model achieves Dice scores of 88.8% for OD and 86% for OC with an IoU of 79.8% and 75.4%. On the DRISHTI-GS dataset, it achieves Dice scores of 80.2% for OD and 87.5% for OC with an IoU of 69.6% and 78.2%. For boundary evaluation, ASPPIAU-Net achieves Hausdorff Distance (HD) of 4.17 for OD and 3.94 for OC on REFUGE and 5.60 for OD and 5,61 for OC on DRISHTI-GS, indicating improved boundary alignment. For clinical consistency, the model achieves an MAE for VCDR of 0.005 on REFUGE and 0.086 on DRISHTI-GS. Overall, ASPPIAU-Net shows robust and balanced performance through multi-scale contextual learning and attention-based feature refinement. The model improves segmentation quality, particularly for the optic cup, and provides a reliable framework for automated glaucoma screening.

KEYWORDS: Disease-screening; glaucoma; health; health risks; image-segmentation

1 Introduction

Deep learning has been widely used in the medical field, including image segmentation. Image segmentation is an image processing technique that segments parts based on certain characteristics from the background [1]. Image segmentation is often used in the medical world to separate the optical part of the retinal image. Optics in retinal images consist of the optic disc and optic cup. The optic disc is an area on the retina that consists of retinal nerve vessels, while the optic cup is a concave area inside the optic

disc. In the medical world, optic segmentation of retinal images can be used in detecting diseases, one of which is glaucoma. Glaucoma is a chronic, neurodegenerative, and irreversible disease [2]. Glaucoma detection can be done by observing abnormalities in the optic cup and optic disc directly. This observation can be done by the optic cup and optic disc segmentation, which separates the optic cup and optic disc from other parts of the retinal image. The manual segmentation process takes a long time and is subjective [3]. Therefore, an automatic approach is needed to segment the optic cup and optic disc using deep-learning image segmentation.

Image segmentation consists of binary segmentation and semantic segmentation. Binary segmentation divides an image into two classes, with a main object as a feature and other objects considered as background, while semantic segmentation is the process of partitioning an image based on the similarity of pixels in a region with other regions by labeling each pixel in the image. In the optic cup and optic disc segmentation, the approach used is semantic segmentation. If the segmentation is done in binary, then the segmentation of each part is done separately, and the segmentation results of each part are combined as the final result. This process can affect the final segmentation result because the model cannot learn and consider the boundary between the optic cup and optic disc directly, which causes the final result to overlap the two optic parts. The semantic segmentation approach is carried out because semantic segmentation can allow the model to learn the boundary between the optic cup and optic disc areas and learn the various sizes between the optic cup and optic disc images, which can reduce the risk of overlap.

The semantic segmentation approach can be implemented by applying a Convolutional Neural Network (CNN), which effectively extracts image features. U-Net is a popular CNN architecture applied to image segmentation [4]. U-Net has a "U" shaped architecture consisting of encoder and decoder parts connected by a bridge [5]. The encoder process gradually reduces feature resolution and extracts image features, while the decoder gradually restores feature resolution and reconstructs features extracted at the encoder stage. The bridge connects the encoder and decoder sections. The U-Net architecture contains numerous feature maps, enabling it to effectively extract and learn features from images [6]. Although the U-Net is popularly used in segmentation, too deep a U-Net network can cause loss of feature details in the feature learning process in the U-Net encoder.

The encoder in U-Net extracts hierarchical features through successive downsampling operations. These operations are commonly implemented using max pooling [6]. While this strategy effectively captures dominant semantic features, repeated spatial reduction may lead to the loss of fine-grained spatial details, which can affect segmentation accuracy. To address this limitation, multi-scale contextual feature extraction is required to enrich feature representation at deeper layers. To address these issues, multi-scale feature learning has been introduced in various studies. Atrous Spatial Pyramid Pooling (ASPP) has been introduced to capture multi-scale contextual information by employing parallel atrous convolutions with different dilation rates. By expanding the receptive field without concurrent spatial downsampling, ASPP facilitates the network's aggregation of broader contextual information while maintaining spatial resolution [7].

On the other hand, decoder U-Net is responsible for restoring spatial resolution via upsampling operations, which are typically implemented using transposed convolution [8]. This mechanism restores spatial dimensions and refines features extracted during the downsampling process [6,8]. However, despite its effectiveness, transposed convolution can introduce reconstruction artefacts that degrade segmentation quality [9]. These artifacts can affect the smoothness and consistency of the predicted boundaries. Moreover, conventional convolution layers applied after upsampling may not sufficiently capture feature variations across multiple spatial scales during reconstruction. To overcome these limitations, it is considered a potential strategy to enhance feature refinement and improve segmentation performance by incorporating a multi-scale feature extraction mechanism in the decoder, such as the Inception module [10].

Inception is a block of Google-Net architecture that uses kernels of different sizes in the convolution layer that can capture features at various scales and directions efficiently [10]. Inception uses different kernels that are 1 × 1 and 3 × 3 in size. The use of different kernels in inception blocks aims to learn more detailed features with different scales and combine the transposed convolution features that can help reconstruct image features in more detail [10]. Adding inception blocks to the decoder can reduce the checkerboard pattern produced by the transposed convolution layer.

The U-Net architecture effectively preserves spatial information through skip connections that transfer feature maps from the encoder to the decoder, thereby improving localization accuracy. These connections help retain fine-grained details that could be lost during downsampling. However, all features from the encoder are passed directly to the decoder without any selection mechanism, which can introduce irrelevant background information [11]. This limitation is more critical in retinal image segmentation because the optic disc and optic cup occupy relatively small regions. To address this, Attention Gates selectively emphasize relevant features and suppress less important information in the skip connections [12]. By guiding the network to focus on relevant regions, attention gates can improve segmentation performance, especially for structures with subtle boundaries, such as the optic cup.

Based on these limitations, this study proposes a novel architecture, namely ASPPIAU-Net, which integrates Atrous Spatial Pyramid Pooling, Inception modules, attention gates, and U-Net into a unified framework to enhance segmentation performance. The ASPP module is incorporated into the bottleneck to enhance multi-scale contextual feature representation. Inception modules are embedded into the decoder to improve feature reconstruction and reduce artifacts caused by transposed convolution. Furthermore, attention gates are applied to the skip connections to selectively emphasize relevant features from the encoder while suppressing irrelevant background information, thereby improving the model's ability to focus on important regions during feature fusion. Additionally, the proposed model simultaneously segments the optic cup and optic disc, allowing for more precise learning of the boundaries between the two structures.

The main contributions of this study are summarized as follows:

1. Proposing a novel hybrid architecture, ASPPIAU-Net, which integrates U-Net, ASPP, Inception, and attention gates into a unified segmentation framework.
2. A dual multi-scale learning strategy is introduced, where ASPP enhances contextual feature extraction at the bottleneck, while Inception modules improve feature reconstruction in the decoder. In addition, attention gates are applied to skip connections to refine feature selection and suppress irrelevant information.
3. Simultaneously performing semantic segmentation of the optic cup and optic disc enables the model to learn inter-structure boundaries more effectively.
4. A comprehensive performance evaluation was conducted using multiple metrics, including accuracy, sensitivity, specificity, Dice score, IoU, and Hausdorff Distance (HD).

The remainder of this paper is organized as follows. Section 2 reviews the related work relevant to this study. Section 3 presents the proposed method in detail, including the model architecture and preprocessing techniques. Section 4 describes the experimental setup. Section 5 provides the results and discussion. Finally, Section 6 concludes the paper by summarizing the main findings and suggesting directions for future work.

2 Related Work

Deep learning, particularly Convolutional Neural Networks (CNNs), has been widely applied for optic cup and optic disc segmentation in retinal images. Among various architectures, U-Net has become

one of the most popular models due to its encoder–decoder structure and strong capability in extracting hierarchical features. Several studies have employed the U-Net for optic segmentation tasks with promising results. Xiao et al. used a U-Net model with a small kernel to segment the optic disc, reporting strong performance with an accuracy of 95.98%, F1-score of 90.45%, and IoU of 94.77%. However, their approach focuses only on optic disc segmentation, without considering the optic cup, which limits its ability to capture the relationship between the two structures [13]. Desiani et al. applied a U-Net-based approach for retinal semantic segmentation and reported performance metrics including accuracy, sensitivity, specificity, and IoU exceeding 90%. However, the evaluation is presented only as overall performance, without explicitly distinguishing between optic disc and optic cup regions [14]. Tadisetty et al. used a U-Net model to segment the optic cup and optic disc, achieving a Dice score above 90% for the optic disc. However, their approach relied on binary segmentation, in which the optic cup and optic disc were segmented separately. Consequently, the Dice score for the optic cup was relatively low, below 65% [4]. Similarly, Joshua et al. implemented the U-Net for binary segmentation and obtained high Dice scores above 95%. Nevertheless, their Intersection over Union (IoU) results were limited, reaching 88% for the optic disc and 79% for the optic cup [15]. These results suggest that binary segmentation approaches may have difficulty accurately capturing the relationship between the optic cup and optic disc.

Several studies have incorporated multi-scale feature extraction techniques into the U-Net to improve segmentation performance. One widely used approach is atrous spatial pyramid pooling (ASPP), which uses parallel atrous convolutions with different dilation rates to capture contextual information at multiple scales while preserving spatial resolution (6). Kedari et al. integrated ASPP into the U-Net and reported precision, recall, and Dice scores below 82% for optic cup and optic disc segmentation [16]. Xia et al. applied ASPP and achieved a high Dice score of 97.18%; however, the intersection over union (IoU) remained suboptimal at 83.42%, indicating that further improvements are needed [17].

In addition to enhancing the encoder, several studies have focused on improving the decoder stage. The decoder usually employs transposed convolution for upsampling, which can result in reconstruction artifacts, such as checkerboard patterns, that affect segmentation quality. To address this issue, the Inception module was proposed as a multi-scale feature extraction mechanism in the decoder. Tulsani et al. incorporated Inception modules into the U-Net architecture, achieving IoU and Dice scores above 90% for both the optic cup and the optic disc. However, their method performed segmentation separately for each structure [18]. Neto et al. proposed a U-Net with Inception for semantic segmentation and achieved Dice scores above 80% for both regions. Nevertheless, the intersection over union (IoU) remained below 75%, and the evaluation metrics used were limited [2]. Desiani et al. used a modified U-Net with Inception blocks in both the encoder and decoder for semantic segmentation. The model achieves high performance. It has accuracy and specificity above 90%. However, the results of IoU remain below 60% [19].

Attention mechanisms have been widely explored as a means of improving feature representation by enabling the model to focus on relevant regions. Xiao et al. [20] introduced an attention-based U-Net model that achieved an accuracy of 90.4%, a precision of 88.6%, and a Dice score of 88.1%. Using attention enhances the model's ability to emphasize important features. However, this method is limited to single-structure segmentation because it only focuses on the optic disc and ignores optic cup segmentation. Although previous studies have demonstrated promising results, several limitations remain. Many approaches still rely on binary segmentation, which limits the model's ability to learn the spatial relationship between the optic cup and optic disc. Furthermore, existing methods often improve only one part of the architecture, either the encoder or the decoder, rather than addressing both simultaneously. These limitations underscore the necessity of a comprehensive approach that incorporates multi-scale feature learning in both the

feature extraction and reconstruction stages to attain more precise and robust segmentation. Comparison of state-of-the-art methods is presented in Table 1.

Table 1: Comparison of state-of-the-art methods in retinal image optical segmentation.

Study	Methodolgy	Advantages	Limitations
Xiao et al. (2025) [20]	Attention based on U-Net	Achieved good performance with an accuracy, precision, and dice above 85%	Limited to a single structure, no OC segmentation
Desiani et al. (2024) [19]	U-Net Inception with residual blocks	Achieved good performance with accuracy and specificity above 90%	IoU below 60%
Xiao et al. (2024) [13]	Simple CNN U-Net	High accuracy (95.98%), strong optic disc segmentation	Only limited to the optic disc, not for the optic cup segmentation
Desiani et al. (2024) [14]	U-Net	High segmentation performance (Acc, Sen, Spe, IoU > 90%)	Evaluation reported only as overall metrics; no separate analysis for optic disc and optic cup
Tadisetty et al. (2023) [4]	U-Net	High Dice score (>90%) for optic disc, showing strong performance for prominent structures	Low Dice for optic cup (<65%); Segmentation performed separately
Kedari et al. (2023) [16]	Modified U-Net with ASPP	Improved contextual information through multiple dilation rates	Overall performance was still below 82%
Xia et al. (2022) [17]	U-Net	Excellent Dice score (97.18%)	IoU remains moderate (83.42%)
Neto et al. (2022) [2]	U-Net Inception	High dice score for both regions (>80%)	IoU below 75% and limited evaluation metrics
Tulsani et al. (2021) [18]	U-Net with inception modules in decoder	High Dice and IoU (>90%) for both OD and OC	Segmentation performed separately
Joshua et al. (2019) [15]	U-Net	Excellent Dice score (>95%)	IoU remains limited (88% OD, 79% OC), binary segmentation

3 Method

The method used in this study consists of data description, data preprocessing, ASPPIAU-Net architecture implementation, and performance evaluation. The flow of methods used in this study can be seen in Fig. 1.

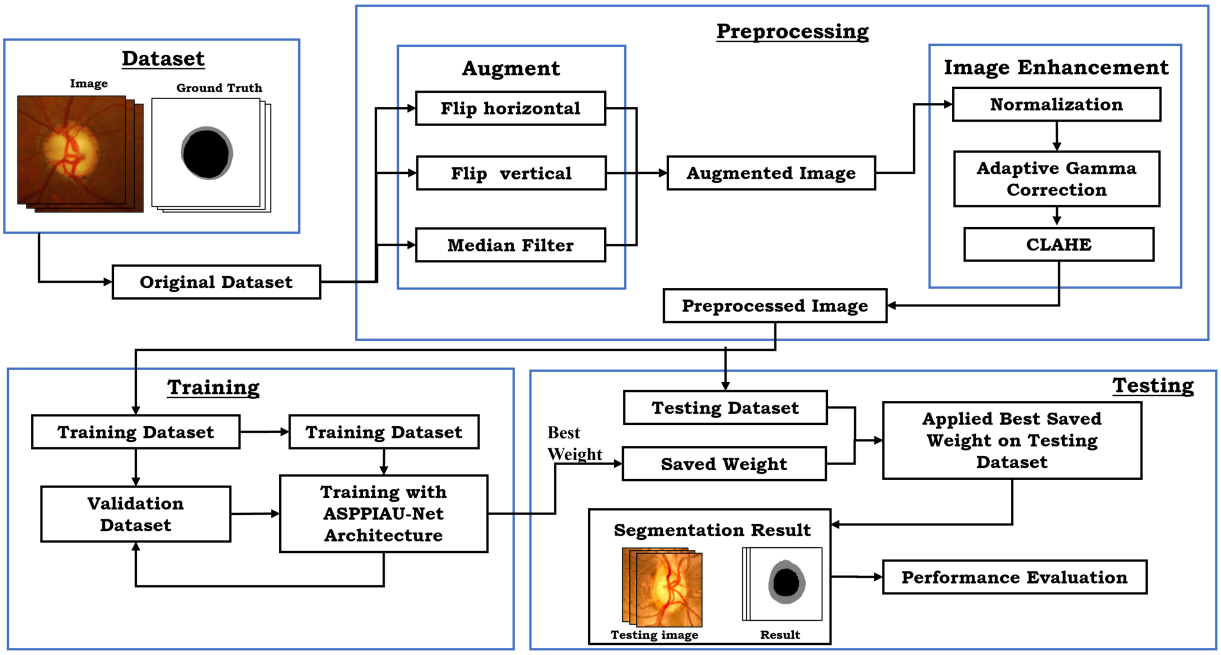

Figure 1: Flowchart of retinal image optical segmentation study methods with ASPPIAU-Net architecture.

3.1 Data Description

In this study, the data used are retinal images from the Drishti-GS dataset and the Refuge dataset, each of which consists of images and ground truth. The Drishti-GS dataset consists of 101 retinal images collected from Aravind Eye Hospital, which were labeled by glaucoma experts. The Refuge dataset consists of 400 images from China, which are labeled by trained ophthalmologists from Zhongshan Ophthalmic Center affiliated with Sun Yat-sen University. The Drishti-GS and Refuge datasets can be accessed on the Zenodo website [21]. The publicly available versions of both datasets are provided in a pre-cropped format with a resolution of 800 × 800 pixels. Each dataset includes annotations for three classes: optic cup, optic disc, and background. Fig. 2 shows examples of the images and corresponding labels used in this study.

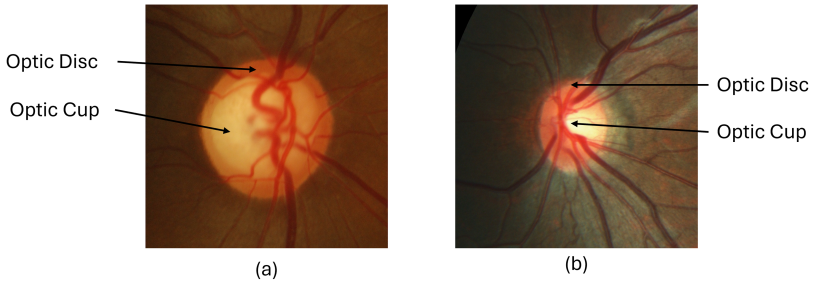

Figure 2: Retinal image and labeled area: (**a**) drishti-GS dataset, (**b**) refuge dataset.

3.2 Data Augmentation

Data augmentation is a technique that increases the amount and diversity of data by modifying existing data. Increasing the diversity of the training data can improve the performance of the model during the evaluation stage. The study used a combination of flipping rotation and median filter techniques. Flipping is a simple augmentation technique that flips images horizontally or vertically, and rotation rotates images by a certain angle to introduce additional geometric variation. A median filter replaces the value of a pixel with the median value of its surrounding pixels. These techniques introduce data variation and reduce noise in the images. Data augmentation in this study is shown in Fig. 3.

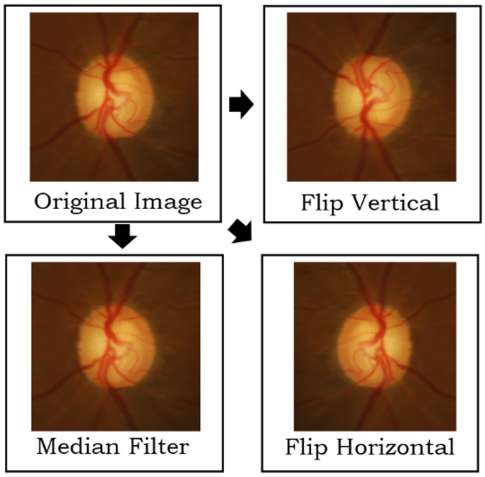

Figure 3: Data augmentation in this study.

3.3 Image Enhancement

This study involved a series of image quality enhancement steps before the primary process. First, gamma correction was applied with a gamma value of 1.2 to adjust pixel intensity and proportionally enhance image contrast [22,23]. Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied with a clip limit of 2 with a grid size of 16 to prevent excessive contrast enhancement while improving local intensity distribution. This process enhanced visual details without causing saturation. The images were resized from 800 × 800 pixels to 128 × 128 pixels to ensure uniform input dimensions for the model and to reduce computational complexity while preserving essential structural information. The final stage involved image normalization to scale pixel intensity values into a uniform range before being used as input to the model. Image normalization was performed by min-max normalization. The image enhancement stages are shown in Fig. 4.

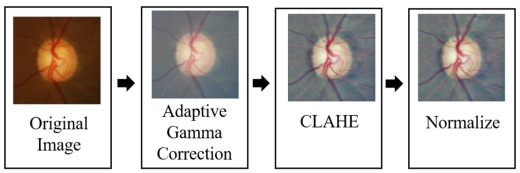

Figure 4: Image enhancement on retinal image segmentation.

3.4 ASPPIAU-Net

The proposed ASPPIAU-Net is a modified U-Net architecture that integrates an ASPP module after the final max pooling operation in the encoder and incorporates an Inception module within each decoder block. The ASPP module is positioned at the bottleneck stage to improve the representation of multi-scale contextual features at the deepest feature level. Employing atrous convolutions with different dilation rates enables the module to aggregate broader contextual information from the pooled feature maps before they enter the bridge layer. The decoder is enhanced by applying an Inception module after each transposed convolution operation. This design aims to refine the reconstructed features and reduce the checkerboard artefacts that are commonly produced by transposed convolutions. Furthermore, Attention Gates are incorporated into the skip connections between the encoder and decoder to selectively filter feature maps. This mechanism allows the network to emphasize relevant regions, such as the optic disc and optic cup, while suppressing irrelevant background information, thereby improving feature fusion and segmentation accuracy. The ASPPIAU-Net architecture can be seen in Fig. 5.

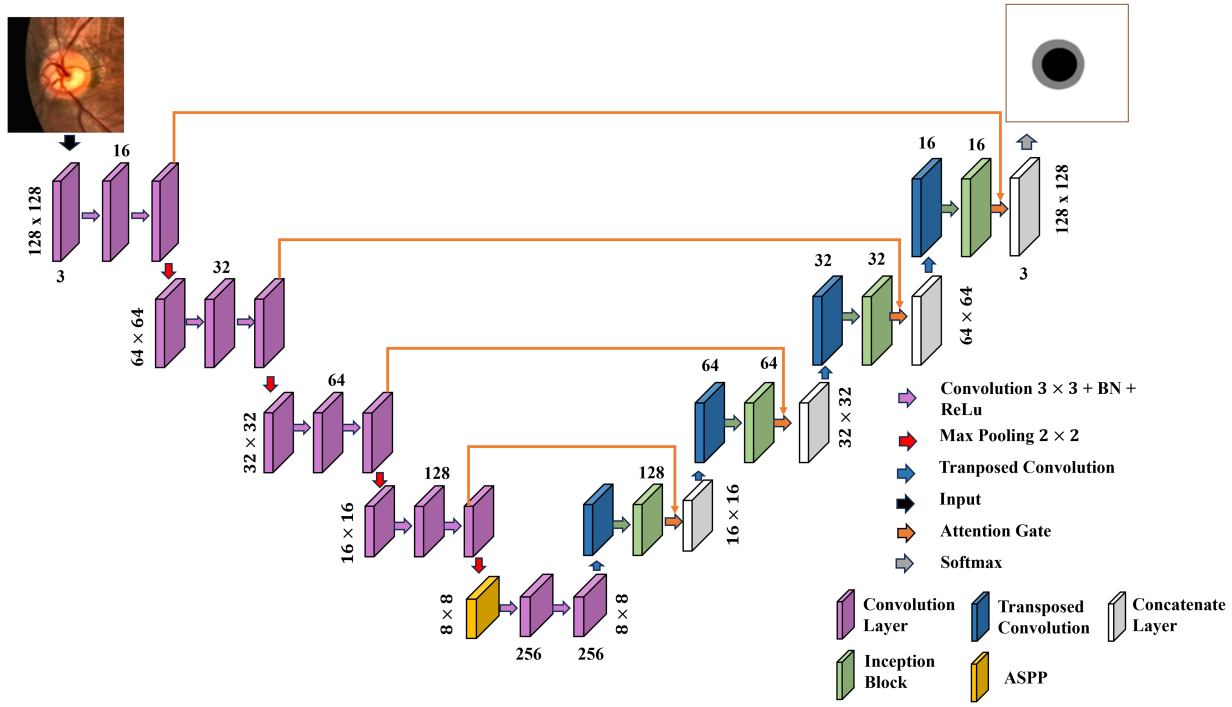

Figure 5: ASPPIAU-Net architecture in this study for retinal image optical segmentation.

ASPPIAU-Net consists of four encoder and decoder blocks connected by a bridge. The input image has a size of 128 × 128 and undergoes feature extraction in the encoder stage. Each encoder block employs convolutional layers with filter sizes of 16, 32, 64, and 128, respectively, followed by batch normalization and ReLU activation. The convolution operation plays an important role in extracting local features from the input image using a stride of 1, as defined in Eq. (1).

$$d_{ij} = \sum_{y=0}^{m-1} \sum_{z=0}^{m-1} \left(t_{y+iq,z+jq} \times k_{y,z} \right) + b_n \tag{1}$$

where d_{ij} is the i-th row and j-th column convolution result matrix entries, where $i = 1, 2, 3, \ldots, m$ and $j = 1, 2, 3..m$, $t_{y+iq,z+jq}$ is the $y + iq$-th row, $z + jq$-th column input matrix entries. $k_{y,z}$ is the y-th row, z-th

column kernel matrix entry. b_n is the bias in the n-th filter. i is the row of the convolution result matrix, j is the column of the convolution result matrix, m is the n-th kernel height, and q is the stride. Furthermore, max pooling with a pool size of 2 × 2 and a stride of 2 is applied at the end of each encoder block. This reduces the spatial dimensions of the feature maps progressively while preserving the most relevant features [6]. After the last encoder layer, the max pooling output is processed by the ASPP module, which captures features at various scales and retains important information from the pooling process. The ASPP block used in the ASPPIAU-Net architecture is shown in Fig. 6.

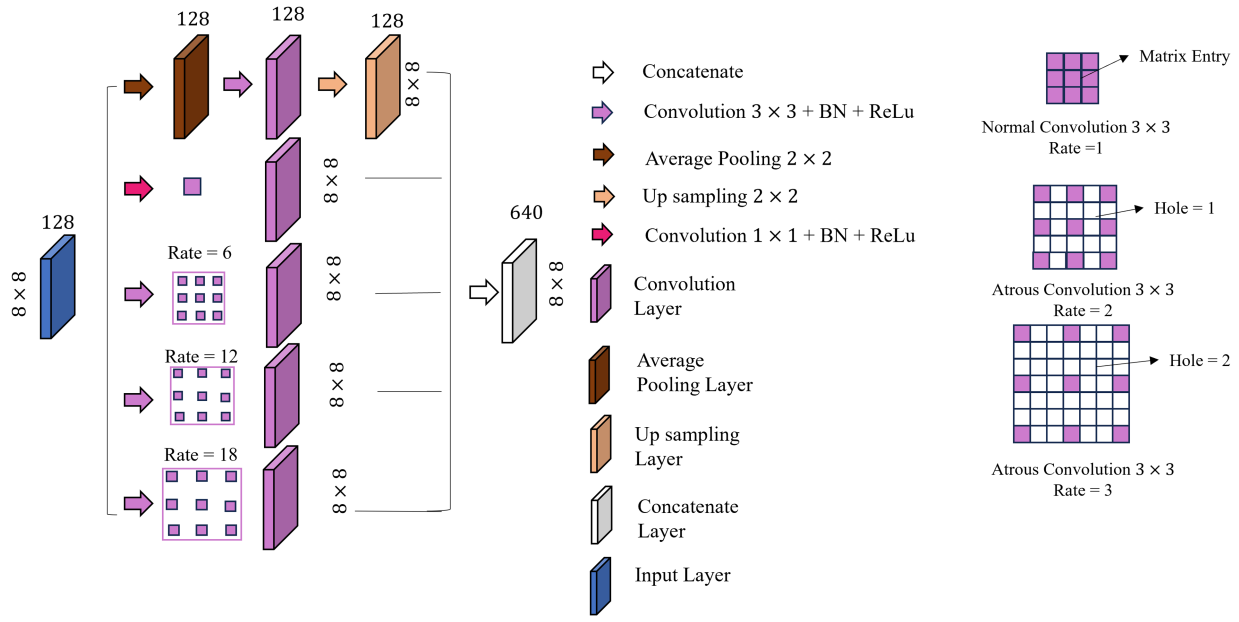

Figure 6: ASPP Block of ASPPIAU-Net architecture in this study for retinal image optical segmentation.

The ASPP module employs five parallel branches. One branch applies global average pooling, followed by a 1 × 1 convolution, batch normalisation, and ReLU activation, and then upsampling. Another branch uses a 1 × 1 convolution. The remaining three branches use atrous convolutions with dilation rates of 6, 12, and 18. The outputs of all the branches are concatenated and undergo further processing using a 3 × 3 convolution with 256 filters, followed by batch normalization and ReLU activation. These features are then forwarded to the bridge layer for spatial upsampling via transposed convolution, after which they enter the decoding stage.

The decoder path progressively reconstructs the segmentation map through four stages, using 128, 64, 32, and 16 filters, respectively. Each stage consists of a transposed convolution for upsampling, followed by convolutional layers with a 3 × 3 kernel, batch normalization, and rectified linear unit (ReLU) activation. Additionally, Inception blocks are incorporated at each decoder stage to enhance multi-scale feature representation. Each Inception block comprises four parallel processing paths, including max pooling followed by a 3 × 3 convolution, 1 × 1 convolution, and 3 × 3 convolution operations. The outputs of all paths are concatenated to form the final feature representation. The number of filters within each Inception block is adapted according to the corresponding decoder stage, following the progression of 128, 64, 32, and 16 filters. Fig. 7 illustrates the general structure of the Inception block used in the proposed architecture.

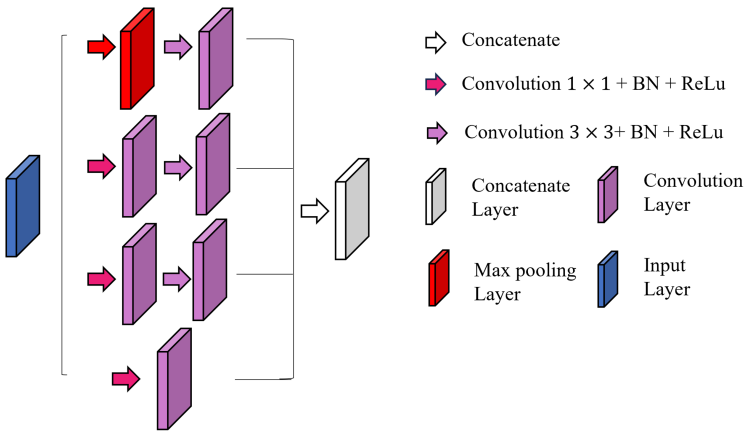

Figure 7: Inception block of ASPPIAU-Net in this study for retinal image optical segmentation.

In this architecture, skip connections are implemented using attention-gated feature concatenation. Specifically, feature maps from the encoder are first refined using an attention gate, which suppresses irrelevant responses and highlights relevant features, before being concatenated with the corresponding decoder features. This mechanism helps preserve important spatial information while reducing noise from irrelevant features. The decoder path progressively reconstructs the segmentation map using convolutional layers with 128, 64, 32, and 16 kernels, respectively. In the final stage, pixel-wise classification is performed using the softmax activation function. This function converts the network's output into probability values for each class, enabling the model to estimate the likelihood that a pixel belongs to a specific class. These probability values are then used to compute the loss during training. To optimize prediction results, this study employs the categorical cross-entropy loss function, which is commonly used for multi-class segmentation tasks. This function measures the difference between the predicted probability distribution and the ground truth labels. The categorical cross-entropy and Dice loss function are given in Eqs. (2) and (3) [24].

$$L_{cc} = -\sum_{i=1}^{N} y_i \log s_i \tag{2}$$

$$L_{DL} = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \tag{3}$$

L_{cc} denotes the result of categorical cross entropy, N denotes the number of classes, y_i denotes the true value of label i, s_i is the probability value for class i generated from the softmax function, L_{DL} denotes the result of the dice loss function, X denotes as segmented ground truth, and Y denotes as predict mask.

4 Experimental Setup

4.1 Dataset Details

This study uses two datasets: Drishti-GS, containing 101 images, and REFUGE, containing 400 images. Each image is annotated with three labels: optic disc, optic cup, and background. The datasets are split into training, validation, and testing sets with a ratio of 64%, 16%, and 20%, respectively. The Drishti-GS dataset was divided into 64 training images, 16 validation images, and 21 testing images, and the REFUGE dataset was divided into 256 training images, 64 validation images, and 80 testing images. The training sets from both datasets were merged into a single training set, and the validation sets were combined. However, the

test sets were kept separate for each dataset to ensure a fair evaluation across different data distributions. Retinal optic images consist of three labels: 0 for the optic disc, 1 for the optic cup, and 2 for the background.

4.2 Implementation Platform

The proposed model was implemented using TensorFlow 2.0.0 with the Keras API. The experiments were conducted on a system equipped with an NVIDIA RTX 3060 GPU, an Intel Core i5 processor @2.90 GHz, and 16 GB RAM on the Windows operating system.

4.3 Training Details

The training procedure is designed to ensure robust model performance and generalization. The dataset is divided into training, validation, and testing sets with a ratio of 64%, 16%, and 20%, respectively. The training and validation sets from both datasets are combined, while the test sets are kept separate to maintain fair evaluation. To improve the model's ability to generalize and reduce overfitting, data augmentation is applied to the training set. The augmentation techniques include gamma correction, hazy augmentation, and geometric transformations such as flipping and rotation. As a result, the number of training samples is increased to 5120 images.

The model was trained for 50 epochs with a batch size of 32, employing the Adam optimizer with a learning rate of 0.001. The model was trained using a combination of Dice loss and categorical cross-entropy, which was selected based on preliminary experiments due to its effectiveness in addressing class imbalance. The dataset was divided into training and validation sets, and the model's performance was monitored by performing validation at the end of each epoch. A data augmentation technique was applied to improve generalization and used early stopping to mitigate overfitting. Model selection was based on the highest validation Dice coefficient, and the weights of the best-performing model were saved. To evaluate model stability, the training process was conducted using multiple random seeds, and all evaluation metrics were recorded for each run. The complete set of training hyperparameters is summarized in Table 2.

Table 2: Training hyperparameters in this study.

Parameter	Value
Optimizer	Adam
Learning Rate	1e−4
Batch Size	32
Epoch	50
Loss Function	Dice + Categorical Cross Entropy
DropOut Rate	0.2
Weight Decay	1e−5
Early Stopping	Patience = 10

4.4 Performance Evaluation

In this study, performance evaluation is carried out using a confusion matrix, which compares the model's predictions with the actual labels in order to assess its overall performance. Several evaluation metrics are then calculated based on this matrix, including accuracy, sensitivity, specificity, dice score, and IoU. Accuracy reflects the overall correctness of the predictions [25]. Sensitivity measures the model's ability to correctly detect object labels, and specificity assesses its ability to identify non-object labels. The dice score

evaluates the balance between sensitivity and specificity. IoU measures the similarity between the predicted segmentation and the ground truth based on the overlap of object boundaries. Hausdorff Distance (HD) measures the maximum distance between the boundary of the predicted segmentation and the ground truth, reflecting the largest boundary error [26,27].

The measurement module utilizes the segmentation results of the optic cup (OC) and optic disc (OD) to calculate the vertical cup-to-disc ratio (VCDR), which serves as a critical quantitative indicator for glaucoma diagnosis. The calculation of CDR is defined in Eq. (4) [28,29].

$$CDR = \frac{V_{cup}}{V_{disc}} \tag{4}$$

where V_{cup} and V_{disc} denote the vertical diameters of the optic cup and optic disc, respectively. A VCDR value greater than 0.6 is commonly considered a potential indicator of glaucoma.

5 Result and Discussion

5.1 Quantitative Performance Evaluation

The proposed model is evaluated on independent testing datasets to assess its generalization capability. The segmentation performance is quantified using the accuracy, sensitivity, specificity, dice score, IoU, and Hausdorff Distance (HD). To provide further insight, additional experiments are conducted by varying architectural components under consistent training settings, ensuring a fair comparison. The results on the Refuge dataset are presented in Table 3.

Table 3: Comparison of performance evaluation of each component on the refuge dataset.

Method	Acc.	Sen (%)		Spe (%)		Dice (%)		IoU		HD	
		OD	OC	OD	OC	OD	OC	OD	OC	OD	OC
U-Net	98	88	87	98.9	99.4	88.2	84.8	78.9	73.6	4.65	4.26
U-Net ASPP	97.9	87.7	85.4	99	99.5	88.2	85.1	79	74.1	4.27	4.09
U-Net Inc	97.9	88.7	86.9	98.8	99.4	88.2	84	78.9	72.5	4.52	4.28
U-Net Att	97.9	87.5	87.2	98.9	99.4	88	84.5	78.6	73.2	4.32	4.14
U-Net ASPP Att	97.9	90	83.5	98.7	99.5	88.3	84.5	79	73.2	4.47	4.16
U-Net Inc Att	98	88.8	85.2	98.9	99.6	88.6	85.6	79.6	74.8	4.43	3.83
U-Net ASPP Inc	97.9	88.6	85.5	98.9	99.5	88.3	85.4	79	74.5	4.30	3.97
U-Net Aspp Inc Att	98.1	87.7	87.9	99.1	99.5	88.8	86	79.8	75.4	4.17	3.94

As shown in Table 3, all of the tested U-Net model variants demonstrated high and relatively stable performance in the optic disc and optic cup segmentation tasks on the Refuge dataset. The baseline U-Net model achieved an accuracy of 98%, with Dice scores of 88% and 87% for the optic disc and optic cup, respectively, as well as an intersection over union (IoU) of 78.9% and 73.6%, respectively. Adding the ASPP, Attention, and Inception modules, either individually or in combination, yielded varying performance improvements across several metrics, particularly for the optic cup, which is more challenging to segment than the optic disc.

The ASPP module generally enhances the model's ability to capture multi-scale context, as evidenced by increased intersection over union (IoU) and reduced Hausdorff distance (HD) values. Meanwhile, the Inception module improves feature representation, leading to an increase in the Dice score, especially for the

optic cup. The model's ability to focus on important areas is enhanced by the Attention module, although the overall improvement is not very significant. Combining these three modules yields the most optimal results. The U-Net ASPP Inception Attention model performed best overall with an accuracy of 98.1%, the highest Dice score of 86% on the optic cup, and an IoU of 75.4%. Additionally, the model achieved a lower HD score of 3.94, indicating improved accuracy at segmentation boundaries. These results demonstrate that integrating ASPP, Inception, and Attention enhances the model's ability to capture complex features, especially in smaller, more challenging optic cup structures compared to optic discs.

Based on the experimental results in Table 4, all of the tested U-Net model variations demonstrated fairly good performance in the optic disc and optic cup segmentation tasks. Accuracy ranged from 95.4% to 96.2%. The baseline U-Net model achieved Dice scores of 78.5% and 87.4% and an intersection over union (IoU) of 64.6% and 77.6%, respectively. This indicates that the baseline performance is already quite stable, particularly for the optic cup compared to the optic disc. The addition of modules such as ASPP, Inception, and Attention had varying effects on model performance. The Inception module generally demonstrated a more consistent contribution than the others, improving the Dice score and IoU, especially for the optic disc. Meanwhile, ASPP provided the capability to capture multi-scale context; however, in some combinations, performance decreased, especially when combined with Attention alone. The attention module itself does not always significantly improve performance and sometimes lowers evaluation metrics, indicating that the feature-focusing mechanism is not yet optimal without robust feature extraction support.

Table 4: Comparison of performance evaluation of each component on DRISHTI dataset.

Method	Acc.	Sen (%)		Spe (%)		Dice (%)		IoU		HD	
		OD	OC	OD	OC	OD	OC	OD	OC	OD	OC
U-Net	96.1	79.9	83.5	97.7	99.2	78.5	87.4	64.6	77.6	5.77	5.78
U-Net ASPP	95.8	78.6	80.4	97.5	99.4	77	86.3	62.6	75.9	6.08	6.19
U-Net Inc	96.2	78.5	83.2	97.9	99.2	78.5	87	68.5	77	5.72	5.79
U-Net Att	95.7	78.6	77.9	97.4	99.5	76.5	85.1	62	74	6.51	6.48
U-Net ASPP Att	95.5	80.6	67	97	99.4	76.6	76.1	62.1	61.4	7.01	6.91
U-Net Inc Att	95.5	76.4	77.9	97.4	99.5	75.4	85.1	60.5	74.1	6.46	6.37
U-Net ASPP Inc	95.4	78	76.2	97.2	99.4	75.6	83.8	60.8	72.1	6.94	6.78
U-Net Aspp Inc Att	96.1	82	80.5	97.6	99.4	80.2	87.5	69.6	78.2	5.60	5.61

In this experiment, the U-Net ASPP Inception Attention model was the best, achieving an accuracy of 96.1%, the highest Dice score of 87.5% for the optic cup, and an IoU of 78.2%. This model also produced the lowest Hausdorff distance values, 5.60 for the optic cup and 5.61 for the optic disc, indicating the most accurate segmentation boundary quality compared to other models. Overall, these results demonstrate that combining the three modules can improve segmentation performance, especially for complex optic cup (OC) structures. However, not all combinations yield consistent improvements across every metric. Despite these improvements, however, the intersection over union (IoU) values remain relatively moderate, particularly for the optic disc (69.6%) and optic cup (78.2%), indicating that the predicted segmentation regions do not fully overlap with the ground truth. This suggests that the model may struggle to accurately capture fine-grained boundaries, especially in regions with low contrast and ambiguity. The sensitivity values support this limitation, showing improvements of 82% for the optic disc and 80.5% for the optic cup. However, these values also indicate that some regions are still missed during segmentation. Therefore, further enhancements,

such as incorporating boundary-aware loss functions or refining feature extraction mechanisms, may be necessary to improve region overlap and segmentation precision.

5.2 Ablation Study on Dilation Rate

To further analyze the impact of architectural configuration, an ablation study is conducted on various dilation rate settings within the ASPP module. This analysis aims to identify the most effective configuration for capturing multi-scale contextual information. An ablation study to evaluate the effect of various dilation rate settings can be seen in Table 5.

Table 5: Ablation study of various dilation rates on ASPPIAU-Net.

Dataset	Dilation Rate	Dice (%)		IoU		HD	
		OD	OC	OD	OC	OD	OC
Refuge	**1, 3, 12, 18**	87.5	82.5	77.9	70.3	4.58	4.32
	1, 6, 12, 24	88.1	84.1	78.8	72.6	4.45	4.29
	1, 2, 4, 8	88.3	84	79.1	72.5	4.92	4.26
	1, 6, 12, 18	88.8	86	79.8	75.4	4.17	3.94
Dhristi-GS	**1, 3, 12, 18**	78.3	87.5	64.3	77.8	5.33	5.47
	1, 6, 12, 24	76.9	84.8	62.6	73.6	6.56	6.60
	1, 2, 4, 8	79.4	86	65.8	75.5	6.17	6.29
	1, 6, 12, 18	80.2	87.5	69.6	78.2	5.60	5.61

An ablation study of various dilation rate configurations in the ASPP module shows that the choice of dilation rates significantly impacts segmentation performance in both datasets. For the Refuge dataset, configurations 1, 6, 12, and 18 achieve the best overall performance with the highest Dice scores of 88.8% on optic disc and 86% on optic cup, and the highest IoU values of 79.8% optic disc and 75.4% optic cup. Additionally, this configuration produces the lowest Hausdorff distance (HD), particularly for the optic cup (3.94), indicating more precise boundary delineation. While other configurations, such as (1, 2, 4, 8), yield competitive Dice and IoU values, they result in higher HD, suggesting less precise boundary segmentation.

A comparable pattern is evident in the Drishti-GS dataset, where the configuration dilation rate 1, 6, 12, 18 once more delivers optimal equilibrium across all assessment metrics. This configuration achieves the highest Dice score for the optic disc (80.2%) and maintains strong performance for the optic cup (87.5%), along with the highest IoU values (69.6% for optic cup and 78.2% for optic disc). Furthermore, it produces lower HD values compared to most other configurations, indicating improved boundary accuracy. In contrast, configurations with smaller dilation rates (1, 2, 4, 8) or larger dilation gaps (1, 6, 12, 24) tend to produce inconsistent results by either missing fine details or failing to capture sufficient contextual information.

Overall, these results suggest that a balanced dilation rate configuration, such as (1, 6, 12, 18), is more effective in capturing multi-scale contextual features without sacrificing spatial detail. This balance is particularly important for segmenting the optic cup, which has more complex spatial boundaries. Therefore, the selected configuration is used as the optimal setting for the proposed model in subsequent experiments.

5.3 Loss Function Analysis

In addition to architectural design, the choice of loss function is crucial for segmentation performance. A comparative analysis of different loss functions was performed to determine the most suitable training strategy for the proposed model, as shown in Table 6.

Table 6: Comparison of performance evaluation of various loss functions on ASPPIAU-Net.

Dataset	Loss Function	Acc	Loss	Dice	IoU
Refuge	CCE	97.9	0.0654	90.76	83.7
	Dice	97.9	**0.0203**	90.99	84.05
	CCE + Dice	**98.1**	0.0601	**91.38**	**84.67**
Dhristi-GS	CCE	**96.2**	0.1232	88.25	80.36
	Dice	95.9	**0.0407**	87.46	78.8
	CCE + Dice	96.1	0.1278	**88.52**	79.95

Note: *Bold values indicate the best results for each evaluation.

A comparison of different loss functions reveals that the choice of loss function significantly impacts segmentation performance across both datasets. On the Refuge dataset, the combination of categorical cross-entropy (CCE) and Dice loss achieves the best overall performance, with the highest accuracy (98.1%), Dice score (91.38%), and intersection over union (IoU) (84.67%). While Dice loss alone yields slightly better overlap metrics (Dice and IoU) than CCE, the combined loss provides more balanced optimization by leveraging both pixel-wise classification and region overlap. This results in superior overall performance. A comparable pattern is evident in the Drishti-GS dataset. The CCE + Dice combination again produces competitive results, achieving the highest Dice score (88.52%) and maintaining stable accuracy (96.1%). While CCE alone slightly outperforms Dice loss in terms of IoU (80.36%), the difference is marginal, and the combined loss demonstrates better consistency across metrics. In contrast, Dice loss alone results in slightly lower performance, particularly in IoU. This indicates that relying solely on region-based optimization may not be sufficient for capturing detailed structures in more challenging datasets.

Overall, these results suggest that combining CCE and Dice loss provides a more robust training strategy, balancing pixel-level accuracy and region overlap. This is particularly beneficial for medical image segmentation tasks, where accurate classification and precise boundary delineation are both essential.

5.4 Training Analysis

To evaluate the training behavior of the proposed model, the learning curves are first analyzed in terms of accuracy and loss. These metrics provide a general overview of model convergence and training stability throughout the training process. The accuracy and loss curves are presented in Fig. 8.

Based on Fig. 8, the model was initially set to train for 50 epochs. However, the training process stopped early, around epoch 25, due to the implementation of early stopping with a patience value of 10. This indicates that the model did not show significant improvement in validation performance beyond this point. In Fig. 8a, training and validation accuracy both increase rapidly during the early epochs and begin to stabilize around epoch 5. Training accuracy continues to improve slightly, reaching nearly 99%, while validation accuracy stabilizes at around 97%–98%. The small gap between the two accuracies suggests that the model generalizes well without significant overfitting. Similarly, the loss curves in Fig. 8b show a consistent decrease in both training and validation loss. Training loss decreases steadily throughout the epochs, while validation loss drops rapidly in the early stages and then stabilizes with minor fluctuations. The absence of

significant divergence between the two losses further indicates that overfitting is well controlled. Overall, the early stopping mechanism effectively prevents the model from training beyond the optimal point, ensuring efficient convergence while maintaining good generalization performance. Additionally, the Dice coefficient and IoU curves that are related to the same experiment are shown in Fig. 9.

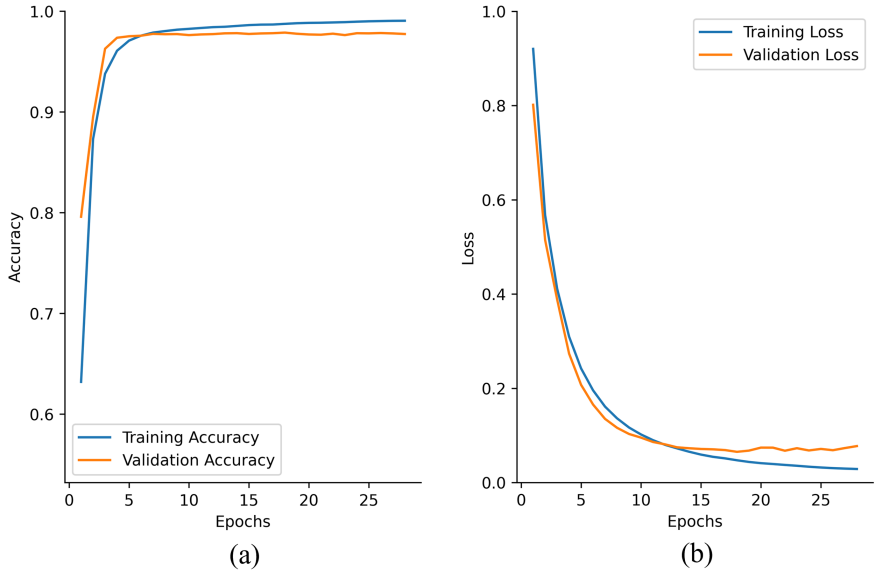

Figure 8: Graphs of (**a**) accuracy and (**b**) loss of segmentation of the optic cup and optic disc in this study.

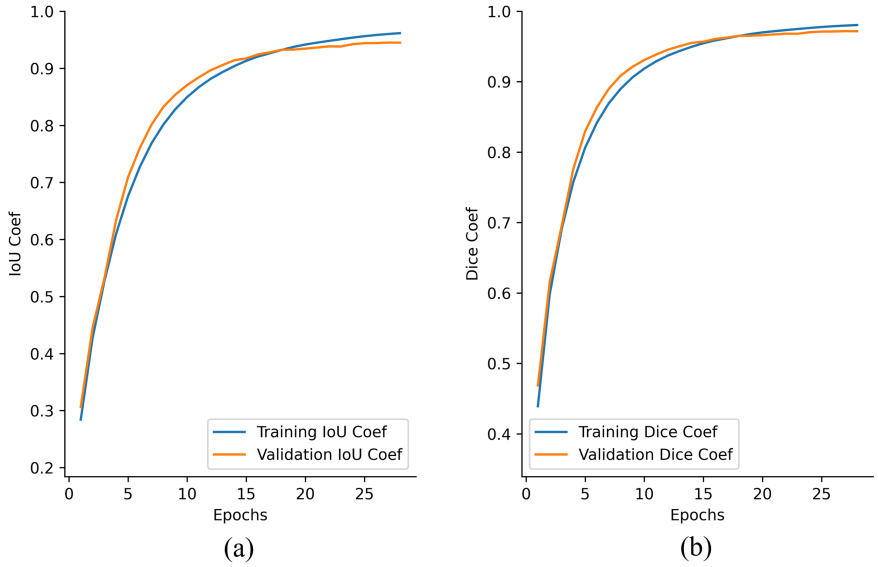

Figure 9: Graphs of (**a**) IoU and (**b**) Dice of segmentation of optic cup and optic disc in this study.

As shown in Fig. 9, both the Dice and IoU curves exhibit a rapid increase during the initial epochs, suggesting that the model swiftly learns to identify the primary structure of the target regions. After epoch 10, improvement becomes more gradual, suggesting the model refines finer details rather than learning new coarse features. Throughout the training process, the training and validation curves for both Dice and IoU remain closely aligned, with only a small gap between them. This indicates good generalization performance

and suggests that the model does not suffer from significant overfitting. Validation Dice stabilizes at a high value (around 0.97), and IoU reaches approximately 0.94–0.95, confirming strong segmentation performance. Overall, these results demonstrate that the model achieves stable convergence and maintains consistent performance across the training and validation sets. This further supports the effectiveness of the proposed approach for accurate segmentation.

5.5 Feature Map Analysis

To further analyze how the model learns hierarchical representations, Fig. 10 visualizes feature maps from selected layers of the encoder, bottleneck, and decoder. For clarity, only a subset of feature maps is presented because visualizing all channels would be impractical due to the high dimensionality of the feature space.

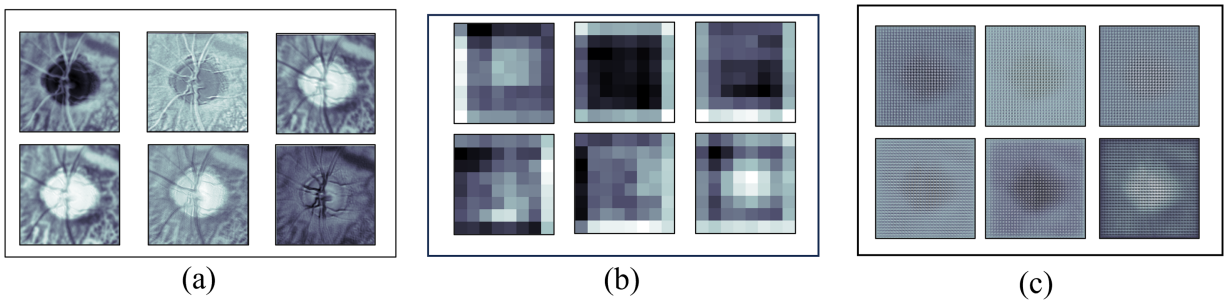

Figure 10: Feature maps of the proposed ASPPIAU-Net model at different stages: (**a**) encoder, (**b**) bottleneck (ASPP), and (**c**) decoder.

As shown in Fig. 10a, the encoder layers primarily capture low-level visual features, such as edges, intensity gradients, and fine vessel structures. The activations preserve spatial similarity to the input image, indicating that the network focuses on local patterns. These features are essential for identifying retinal boundaries and subtle variations in vessel thickness. This information serves as the basis for learning higher-level features. Fig. 10b shows that the feature maps at the bottleneck stage exhibit more abstract and compressed representations. The incorporation of the Atrous Spatial Pyramid Pooling (ASPP) module enables the model to capture multi-scale contextual information by applying dilated convolutions with different receptive fields. Consequently, the feature maps no longer resemble the original image; rather, they encode semantic information related to larger anatomical structures, such as the optic disc and optic cup regions. This multi-scale aggregation allows the network to balance local details and global context effectively, which is essential for accurately segmenting retinal images.

As shown in Fig. 10c, the decoder layers progressively reconstruct spatial information from the encoded features. The feature maps become smoother and more structured, indicating refinement of segmentation-relevant regions. Through successive upsampling and convolution operations, the decoder integrates low- and high-level features, enabling the model to recover fine-grained details, such as the delineation of the boundary between the optic disc and cup. This reconstruction process plays a critical role in improving the precision and continuity of the segmentation output. Overall, these visualizations confirm that the proposed ASPPIAU-Net effectively learns hierarchical and multi-scale representations, contributing to its superior segmentation performance.

5.6 Qualitative Segmentation Results

The segmentation results for each architecture in the ablation study can be seen in Fig. 11.

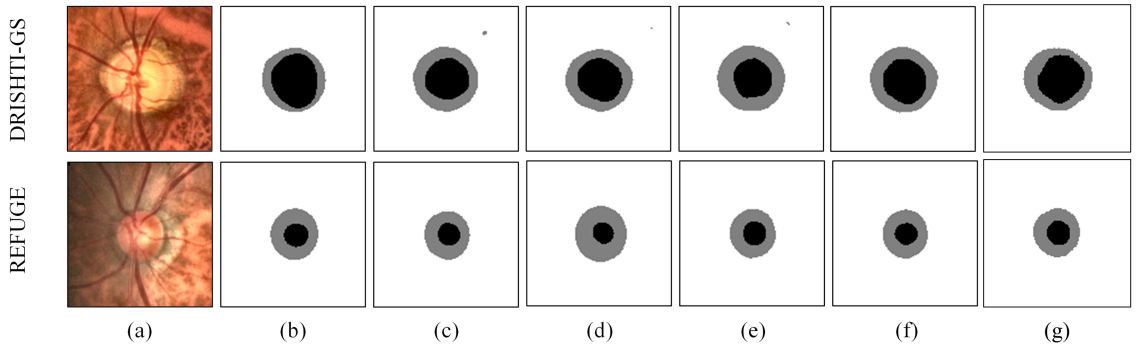

Figure 11: Sample of (**a**) original image, (**b**) ground truth, (**c**) baseline U-Net, (**d**) U-Net ASPP, (**e**) U-Net Inception, (**f**) U-Net ASPP Inception, (**g**) ASPPIAU-Net.

Based on Fig. 11. The baseline U-Net generally produced segmentation results that followed the ground truth structure. However, there were differences in the thickness of the boundaries and the size of the cup area in some parts. U-Net + ASPP showed no significant improvement over the baseline model. Minor artifacts were still visible in some areas, and the boundary improvements were inconsistent. As a result, adding ASPP alone did not lead to noticeable visual changes. U-Net Inception shows a more stable optic disc and cup shape in maintaining proportions, although the precision of the cup boundary is not yet fully consistent with the ground truth. ASPPIAU-Net produced the segmentation closest to the ground truth. The boundaries look smoother and more proportional, with less over-segmentation and under-segmentation. Overall, combining ASPP, Inception, and attention gates provides more stable results than using separate modules or the U-Net baseline.

5.7 ROC Analysis

To provide a more comprehensive evaluation of the model's performance, particularly in terms of class separability, ROC curve analysis is performed. The ROC curves for the three classes are presented in Fig. 12.

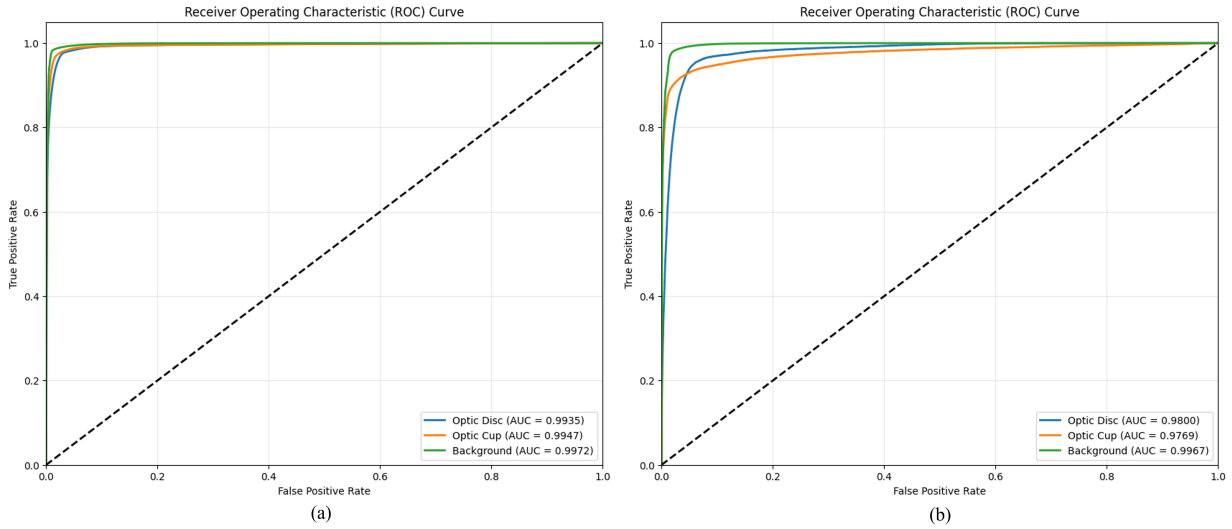

Figure 12: The ROC graphs of ASPPIAU-Net on optic disc and optic cup segmentation: (**a**) Refuge dataset, (**b**) Drishti-GS dataset.

Based on Fig. 12, the curves are located close to the top-left corner. This indicates excellent classification performance across all classes. The model achieves high true positive rates (TPR) at very low false positive rates (FPR), demonstrating strong discriminative capability. The optic cup class attains the highest Area Under the Curve (AUC) values, suggesting that the model effectively identifies this relatively small, low-contrast region. The background class also shows consistently high AUC values, reflecting the model's ability to distinguish foreground structures from irrelevant regions. Meanwhile, the optic disc class exhibits slightly lower AUC values compared to the other classes, possibly due to similarity in intensity and boundary overlap with surrounding retinal structures. The consistent shape of the ROC curves across different experimental settings indicates the robustness and stability of the proposed model. Overall, these results confirm that the ASPPIAU-Net model provides strong classification performance, which supports the earlier quantitative evaluation.

5.8 CDR Analysis

A quantitative analysis was conducted to evaluate the model's ability to estimate the vertical cup-to-disc ratio (VCDR) by comparing the ground truth values (VCDR_GT) with the model's predicted results (VCDR_Pred). The absolute difference between the two was calculated as the absolute error, representing the degree to which the prediction deviates from the reference value. The evaluation was performed on two different test datasets to assess the consistency of the model's performance. The results of the absolute error calculations for the Refuge and Dhristi-GS datasets are presented in Tables 7 and 8.

Table 7: Absolute error of VCDR calculation on the refuge dataset.

Images	Vcdr_GT	Vcdr_Pred	Absolute Error
Img_0.png	0.54	0.55	0.012
Img_1.png	0.562	0.411	0.150
Img_2.png	0.847	0.740	0.106
⋮	⋮	⋮	⋮
Img_78.png	0.387	0.408	0.0204
Img_79.png	0.45	0.545	0.092
	Mean ± std.		0.05 ± 0.04

Table 8: Absolute error of VCDR calculation on the Drishti-GS dataset.

Images	Vcdr_GT	Vcdr_Pred	Absolute Error
Img_0.png	0.538	0.596	0.057
Img_1.png	0.7	0.803	0.103
Img_2.png	0.706	0.660	0.046
⋮	⋮	⋮	⋮
Img_19.png	0.714	0.508	0.205
Img_20.png	0.736	0.71	0.025
	Mean ± std.		0.086 ± 0.056

Based on Table 7, the absolute error values show relatively small variations across images in the Refuge dataset. Some samples exhibit very low errors, indicating that the model's predictions are close to the ground truth. However, there are also cases with higher deviations. Overall, an MAE of 0.05 ± 0.04 was obtained. This

low mean value suggests that the model is accurate in its estimation of VCDR on this dataset. Meanwhile, the small standard deviation indicates that the model's performance is stable and does not vary significantly across samples. Based on Table 8, the absolute error values in the Drishti-GS dataset tend to vary more than in the Refuge dataset. The evaluation results show an MAE of 0.086 ± 0.056. Compared to the Refuge dataset, the higher mean error value indicates that the model's prediction error rate has increased in this dataset. Additionally, the larger standard deviation suggests greater variation in the model's performance across samples, potentially influenced by image complexity or differences in data distribution.

5.9 Discussion

This study proposes a combination of U-Net architecture with ASPP, Inception, and attention gates, namely ASPPIAU-Net, for optic cup and optic disc segmentation in retinal images. The performance results using the ASPPIAU-Net architecture are compared with architectures in other studies in the optic cup and optic disc segmentation in retinal images. A comparison of the performance results of ASPPIAU-Net and other architectures on the Refuge Dataset can be seen in Table 9.

Table 9: Optic cup and optic disc segmentation results with ASPPIAU-Net architecture on the refuge dataset.

Method	Dice		IoU	
	OD	OC	OD	OC
FCN [30]	91.3	84.8	84.03	73.6
EO-Protoseg [31]	–	–	88.21	73.70
U-Net [32]	89.53	88.7	–	–
ASPPIAU-Net	88.8	86	79.8	75.4

As shown in Table 9, ASPPIAU-Net performs competitively in segmenting the optic disc and optic cup (OC) on the REFUGE dataset. The model achieves Dice scores of 88.8 and 86.0 for the OD and OC, respectively, with corresponding intersection over union (IoU) values of 79.8 and 75.4. Compared to FCN [30], ASPPIAU-Net shows lower OD performance in both Dice and IoU. However, it achieves competitive results on OC segmentation. This indicates an improved capability in capturing finer and more localized structures. Examples of these structures include the optic cup. Similarly, ASPPIAU-Net exhibits lower performance in optic disc segmentation when compared to EO-Protoseg [31], suggesting limitations in modeling global structural consistency for larger anatomical regions. When compared to U-Net [32], ASPPIAU-Net does not outperform the baseline consistently across all metrics. U-Net achieves higher OD Dice performance, but ASPPIAU-Net shows competitive or slightly lower performance overall. This indicates that the integration of Atrous Spatial Pyramid Pooling (ASPP) does not uniformly enhance segmentation accuracy across both anatomical structures. However, ASPPIAU-Net maintains relatively stable performance in optic cup segmentation. This suggests that multi-scale feature extraction contributes more effectively to localized structure representation than to global boundary refinement. In general, ASPPIAU-Net's performance is comparable to existing methods on the REFUGE dataset, but it doesn't demonstrate significant superiority. These results highlight the need for further refinement in boundary-aware feature learning to enhance global structural consistency while maintaining strong local feature representation. This would improve the model's overall robustness for clinical retinal image analysis. A comparison of performance results on the Dristhi dataset can be seen in Table 10.

As shown in Table 10, ASPPIAU-Net performs competitively on the Drishti-GS dataset, especially for optic cup segmentation. The model achieves Dice scores of 80.2 and 87.5 for the optic disc and optic cup,

respectively, with corresponding intersection over union (IoU) values of 69.6 and 78.2. While ASPPIAU-Net shows lower OD performance compared to BEAC-Net [33], it has a higher OC Dice score. This indicates that the model is more effective in capturing finer, more localized anatomical structures, particularly the optic cup. Its strong performance in optic cup segmentation suggests that multi-scale feature extraction via Atrous Spatial Pyramid Pooling (ASPP) improves sensitivity to small or complex regions by enhancing contextual representation. However, ASPPIAU-Net achieves lower IoU for optic disc segmentation when compared to EO-Protoseg [31], indicating limitations in capturing precise global boundaries of larger structures. Similarly, compared to DeepLabV3 [32], performance differences are inconsistent across OD and OC metrics, suggesting that architectural enhancements do not uniformly translate into improved segmentation across all anatomical structures. ASPPIAU-Net shows substantially lower performance than SegFormer across all metrics, particularly in Dice and IoU scores for OD and OC [34]. This suggests that transformer-based architectures, like SegFormer, are more effective at modeling long-range dependencies and global contextual information. This leads to more accurate and balanced segmentation results. Overall, the results reveal a clear imbalance in performance between optic disc and optic cup segmentation. ASPPIAU-Net performs relatively better on optic cup structures. This suggests that, although ASPP improves multi-scale contextual learning, it has a greater impact on smaller or more complex regions. Global boundary modeling for larger structures, such as the optic disc, remains challenging. Despite these limitations, ASPPIAU-Net demonstrates competitive performance among CNN-based methods, especially for optic cup segmentation. Further improvements are needed to refine global boundaries and achieve more balanced performance across both anatomical structures.

Table 10: Optic cup and optic disc segmentation results with ASPPIAU-Net architecture on the Dristhi dataset.

Method	Dice		IoU	
	OD	OC	OD	OC
BEAC-Net [33]	86.14	80.8	83.85	76.33
EO-Protoseg [31]	–	–	86.80	71.78
Deeplab V3 [32]	77.86	84.34	–	–
Segformer [34]	90	92	82	86
ASPPIAU-Net	80.2	87.5	69.6	78.2

6 Conclusions

This study proposes ASPPIAU-Net, a deep learning architecture for optic disc (OD) and optic cup (OC) segmentation that integrates multi-scale contextual learning within an encoder–decoder framework. The method is designed to address challenges in retinal fundus image segmentation, particularly boundary ambiguity and anatomical variability, which directly affect the reliability of glaucoma-related measurements such as the vertical cup-to-disc ratio. Experimental results on the REFUGE and DRISHTI-GS datasets demonstrate that ASPPIAU-Net delivers competitive and consistent performance in terms of accuracy, sensitivity, specificity, Dice score, intersection over union (IoU), and Hausdorff distance (HD). HD validation further supports the model's boundary-level precision, which is essential for reliable clinical segmentation. On the REFUGE dataset, the model demonstrates consistent OD and OC segmentation performance, with improved structural representation and boundary coherence. Multi-scale contextual features enhance the model's ability to capture global context and fine boundary details, resulting in more reliable overlap and reduced boundary deviation, as reflected by HD. Similarly, on the DRISHTI-GS dataset, ASPPIAU-Net demonstrates strong generalization ability by maintaining balanced performance across both structures.

The model performs particularly well in optic cup segmentation; multi-scale features contribute to the improved delineation of small, complex regions. However, improvements are not uniform, especially for optic disc segmentation, where achieving global consistency remains challenging. Overall, ASPPIAU-Net provides a robust, clinically relevant segmentation framework that effectively balances region-level accuracy and boundary precision. Future work should focus on enhancing global boundary consistency and computational efficiency to support large-scale clinical deployment.

Acknowledgement: The authors would like to express their gratitude to the Directorate of Research and Community Service, Directorate General of Research and Development, Ministry of Higher Education, Science, and Technology, as well as the Rector of the University through LPPM, for their support and facilitation of this research.

Funding Statement: This research was funded by the Directorate of Research and Community Service, Directorate General of Research and Development, Ministry of Higher Education, Science, and Technology under contract number 109/C3/DT.05.00/PL/2025, and supported by the Rector of the University through LPPM under contract number 0042.034/UN9/SB3.LPPM.PT/2025.

Author Contributions: Anita Desiani: conceptualization, software development, validation, writing—original draft preparation, visualization, supervision, and project administration; Indri Ramayanti: conceptualization, validation, formal analysis, data curation, writing—review and editing, and supervision; Sigit Priyanta: methodology, validation, formal analysis, writing—review and editing, and visualization; Bambang Suprihatin: conceptualization, writing—original draft preparation, and supervision; Muhammad Arhami: methodology, formal analysis, writing—review and editing, and visualization; Deshinta Arrova Dewi: methodology, software development, validation, and data curation; Puspa Sari: software development, writing—review and editing, and visualization. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data used in this study are publicly available and can be accessed at https://doi.org/10.5281/zenodo.8009107.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Surono S, Rivaldi M, Dewi DA, Irsalinda N. New approach to image segmentation: U-Net convolutional network for multiresolution CT image lung segmentation. Emerg Sci J. 2023;7(2):498–506. doi:10.28991/esj-2023-07-02-014.
2. Neto A, Camera J, Oliveira S, Cláudia A, Cunha A. Optic disc and cup segmentations for glaucoma assessment using cup-to-disc ratio. Procedia Comput Sci. 2022;196(2):485–92. doi:10.1016/j.procs.2021.12.040.
3. Juneja M, Singh S, Agarwal N, Bali S, Gupta S, Thakur N, et al. Automated detection of Glaucoma using deep learning convolution network (G-net). Multimed Tools Appl. 2020;79(21):15531–53. doi:10.1007/s11042-019-7460-4.
4. Tadisetty S, Chodavarapu R, Jin R, Clements RJ, Yu M. Identifying the edges of the optic cup and the optic disc in glaucoma patients by segmentation. Sensors. 2023;23(10):4668. doi:10.3390/s23104668.
5. Gornale SS, Kamat P, Hiremath PS, Kumar S, Goh KW. Automated segmentation and trimester-based classification of fetal head circumference in ultrasound images using deep learning techniques. Int J Patt Recogn Artif Intell. 2026;40(4):2552038. doi:10.1142/s021800142552038x.
6. Du G, Cao X, Liang J, Chen X, Zhan Y. Medical image segmentation based on U-Net: a review. JIST. 2020;64(2):20508–1–020508–12. doi:10.2352/j.imagingsci.technol.2020.64.2.020508.

7. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell. 2018;40(4):834–48. doi:10.1109/TPAMI.2017.2699184.
8. Chen Y, Xia R, Yang K, Zou K. MICU: image super-resolution via multi-level information compensation and U-Net. Expert Syst Appl. 2024;245(12):123111. doi:10.1016/j.eswa.2023.123111.
9. Sanjar K, Bekhzod O, Kim J, Kim J, Paul A, Kim J. Improved U-Net: fully convolutional network model for skin-lesion segmentation. Appl Sci. 2020;10(10):3658. doi:10.3390/app10103658.
10. Bhattacharya P, Zolzer U. Convolutional neural network with inception blocks for image compression artifact reduction. In: Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN); 2020 Jul 19–24; Glasgow, UK. doi:10.1109/ijcnn48605.2020.9206676.
11. Kamath A, Willmann J, Andratschke N, Reyes M. The impact of U-Net architecture choices and skip connections on the robustness of segmentation across texture variations. Comput Biol Med. 2025;197(7):111056. doi:10.1016/j.compbiomed.2025.111056.
12. Saifullah S, Dreżewski R, Yudhana A, Wielgosz M, Caesarendra W. Modified U-Net with attention gate for enhanced automated brain tumor segmentation. Neural Comput Appl. 2025;37(7):5521–58. doi:10.1007/s00521-024-10919-3.
13. Xiao Y, Zhao J, Yu Y, Ding X, Liu S, Bao W, et al. SimpleCNN-UNet: an optic disc image segmentation network based on efficient small-kernel convolutions. Expert Syst Appl. 2024;256(4):124935. doi:10.1016/j.eswa.2024.124935.
14. Desiani A, Priyanta S, Ramayanti I, Suprihatin B, Al-Filambany MG, Salamah F. Improved U-Net performance with augmentation for retinal optic segmentation. In: Proceedings of the 2023 International Conference on Informatics, Multimedia, Cyber and Informations System (ICIMCIS); 2023 Nov 7–8; Jakarta Selatan, Indonesia. doi:10.1109/ICIMCIS60089.2023.10348984.
15. Joshua AO, Nelwamondo FV, Mabuza-Hocquet G. Segmentation of optic cup and disc for diagnosis of glaucoma on retinal fundus images. In: Proceedings of the 2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA); 2019 Jan 28–30; Bloemfontein, South Africa. doi:10.1109/robomech.2019.8704727.
16. Kedari B, Kamath R, Arra A, Savitha G, Girisha S. Semantic segmentation of optic disc and optic cup using deep learning. In: Proceedings of the 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT); 2023 Jul 6–8; Delhi, India. doi:10.1109/ICCCNT56998.2023.10308314.
17. Xia X, Huang Z, Huang Z, Shu L, Li L. A CNN-transformer hybrid network for joint optic cup and optic disc segmentation in fundus images. In: Proceedings of the 2022 International Conference on Computer Engineering and Artificial Intelligence (ICCEAI); 2022 Jul 22–24; Shijiazhuang, China. doi:10.1109/ICCEAI55464.2022.00106.
18. Tulsani A, Kumar P, Pathan S. Automated segmentation of optic disc and optic cup for glaucoma assessment using improved UNET++ architecture. Biocybern Biomed Eng. 2021;41(2):819–32. doi:10.1016/j.bbe.2021.05.011.
19. Desiani A, Andriani Y, Ramayanti I, Priyanta S, Suprihatin B, Apriyani CN, et al. Rib-net as modification of CNN architecture for semantic segmentation of optic disc and optic cup. Biomed Eng Appl Basis Commun. 2024;36(6):2450036. doi:10.4015/s1016237224500364.
20. Xiao Y, Ding X, Liu S, Ma Y, Zhang T, Xiang Z, et al. Fusion-attention diagnosis network (FADNet): an end-to-end framework for optic disc segmentation and ocular disease classification. Inf Fusion. 2025;124(8):103333. doi:10.1016/j.inffus.2025.103333.
21. Chen Z, Pan Y, Ye Y, Cui H, Xia Y. A fundus image dataset for domain generalization in joint segmentation of optic disc and optic cup. Zenodo. 2023. doi:10.5281/zenodo.8009107.
22. Desiani A, Adrezo M, Chika Marselina N, Arhami M, Salsabila A, Gibran Al-Filambany M. A combination of image enhancement and U-Net architecture for segmentation in identifying brain tumors on CT-SCAN images. In: Proceedings of the 2022 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS); 2022 Nov 16–17; Jakarta, Indonesia. doi:10.1109/ICIMCIS56303.2022.10017519.
23. Maiyanti SI, Desiani A, Lamin S, Puspitashati, Arhami M, Gofar N, et al. Rotation-gamma correction augmentation on CNN-dense block for soil image classification. Appl Comput Sci. 2023;19(3):96–115. doi:10.35784/acs-2023-27.

24. Yang J, Lu Y, Zhang Z, Wei J, Shang J, Wei C, et al. A deep learning method coupling a channel attention mechanism and weighted dice loss function for water extraction in the Yellow River Basin. Water. 2025;17(4):478. doi:10.3390/w17040478.
25. Rachmad A, Syarief M, Hutagalung J, Hernawati S, Rochman EMS, Asmara YP. Comparison of CNN architectures for *Mycobacterium tuberculosis* classification in sputum images. Ingénierie Des Systèmes D Inf. 2024;29(1):49–56. doi:10.18280/isi.290106.
26. Desiani A, Erwin, Suprihatin B, Riana D, Arhami M, Ramayanti I, et al. Denoised non-local means with BDDU-net architecture for robust retinal blood vessel segmentation. Int J Patt Recogn Artif Intell. 2023;37(16):2357016. doi:10.1142/s0218001423570161.
27. Triasari WBE. CS-based lung covid-affected X-ray image disorders classification using convolutional neural network. J Appl Data Sci. 2024;5(4):1939–48. doi:10.47738/jads.v5i4.371.
28. Virbukaitė S. Methodology for applying deep learning algorithms for glaucoma identification [dissertation]. Vilnius, Lithuania: Vilnius University; 2025.
29. Lv X, Yang Y, Wan C, Zhao J, Chi W, Yang W. Automated cup-to-disc ratio quantification via color fundus photography for chronic glaucoma screening. BMC Med Imag. 2025;25(1):429. doi:10.1186/s12880-025-01981-x.
30. Jin Z. Comparison of fully convolutional networks and U-Net for optic disc and optic cup segmentation. ITM Web Conf. 2025;70(10):03022. doi:10.1051/itmconf/20257003022.
31. Zumarsyah PA, Ardiyanto I, Nugroho HA. Meta-learners for few-shot weakly-supervised optic disc and cup segmentation on fundus images. Comput Biol Med. 2026;201(13):111384. doi:10.1016/j.compbiomed.2025.111384.
32. Liu X, Ma Q, Wang J, Liu X, Zhang Q, Yao J, et al. OD/OC-semantic FPN: an enhanced optic cup and disc segmentation model in color fundus images using improved MaxViT and semantic feature pyramid network. Electron Res Arch. 2025;33(9):5496–517. doi:10.3934/era.2025246.
33. Jiang L, Tang X, You S, Liu S, Ji Y. BEAC-net: boundary-enhanced adaptive context network for optic disk and optic cup segmentation. Appl Sci. 2023;13(18):10244. doi:10.3390/app131810244.
34. Salamah F, Erwin E, Desiani A. Improving optic disc and optic cup segmentation with flip-gamma augmentation and SegFormer. SinkrOn. 2026;10(2):1157–68. doi:10.33395/sinkron.v10i2.15996.