SegImgNet: Segmentation-Guided Dual-Branch Network for Retinal Disease Diagnoses

Xinwei Luo[1], **Songlin Zhao**[1], **Yun Zong**[2], **Yong Chen**[3], **Gui-Shuang Ying**[3], **Lifang He**[1]
[1]Lehigh University, Bethlehem, PA, USA
[2]Guilin University of Electronic Technology, Guilin, Guangxi, China
[3]University of Pennsylvania, Philadelphia, PA, USA
xil620@lehigh.edu, soz223@lehigh.edu, franciszong145@gmail.com, ychen123@pennmedicine.upenn.edu, gsying@pennmedicine.upenn.edu, lih319@lehigh.edu

Abstract

Retinal image plays a crucial role in diagnosing various diseases, as retinal structures provide essential diagnostic information. However, effectively capturing structural features while integrating them with contextual information from retinal images remains a challenge. In this work, we propose segmentation-guided dual-branch network for retinal disease diagnosis using retinal images and their segmentation maps, named SegImgNet. SegImgNet incorporates a segmentation module to generate multi-scale retinal structural feature maps from retinal images. The classification module employs two encoders to independently extract features from segmented images and retinal images for disease classification. To further enhance feature extraction, we introduce the Segmentation-Guided Attention (SGA) block, which leverages feature maps from the segmentation module to refine the classification process. We evaluate SegImgNet on the public AIROGS dataset and the private e-ROP dataset. Experimental results demonstrate that SegImgNet consistently outperforms existing methods, underscoring its effectiveness in retinal disease diagnosis.

Code — https://github.com/hawk-sudo/SegImgNet

Introduction

Retinal imaging, particularly fundus photography, is a non-invasive technique widely used in ophthalmology to capture detailed visualizations of retinal structures. By analyzing these images, clinicians can diagnose not only ocular diseases but also systemic conditions such as hypertension and diabetes (Li et al. 2023; Tan et al. 2024). However, manual interpretation by ophthalmologists is costly, time-consuming, and subject to variability, potentially leading to delays in patient care and inconsistent diagnoses. Therefore, there is an urgent need for automated tools to improve disease detection efficiency through retinal image analysis.

Deep learning has emerged as a promising tool for automating disease detection using retinal images (Zhou et al. 2023; Huang et al. 2023; Zhao et al. 2023). These methods typically leverage established computer vision architectures and employ transfer learning to adapt them for various medical applications, as illustrated in Figure 1(a). For example, RETFound (Zhou et al. 2023), built on the Vision Transformer (ViT) architecture, is pretrained on large-scale datasets and later fine-tuned on retinal image datasets for disease detection. However, despite their effectiveness, these approaches focus primarily on modeling the overall data distribution of retinal images rather than on highlighting structural features of the retina. Critical diagnostic features are often embedded in the fine-grained structural details of the retina elements that may not significantly impact the overall data distribution but are essential for accurate disease diagnosis. Consequently, compared to natural image classification tasks, retinal disease diagnosis requires models with a stronger ability to capture and interpret key structural features. To address this challenge, a common strategy is to segment key retinal structures from retinal images (Li and Liu 2022; Almeida et al. 2024; Wang et al. 2021a; Sivapriya et al. 2024). By isolating diagnostically significant structures, the model can focus on extracting relevant features, as shown in Figure 1(b). For example, (Almeida et al. 2024) utilizes a customized image processing technique to segment retinal blood vessels and feed them into DenseNet121 for disease classification, while (Sivapriya et al. 2024) employs ResEAD2Net for blood vessel segmentation and sub-

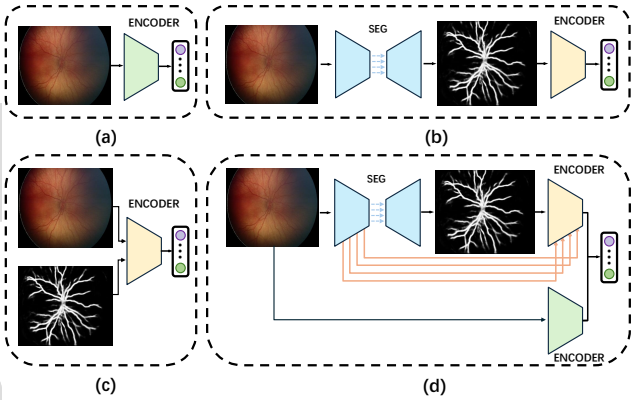

Figure 1: Comparison of different approaches: a) Direct classification of raw images using a standard deep learning model. b) Classification based on the segmented image extracted from a segmentation model. c) Combined classification using both raw and segmented images in a shared encoder. d) Classification using both raw images and enhanced segmentation maps in dual encoders (**ours**).

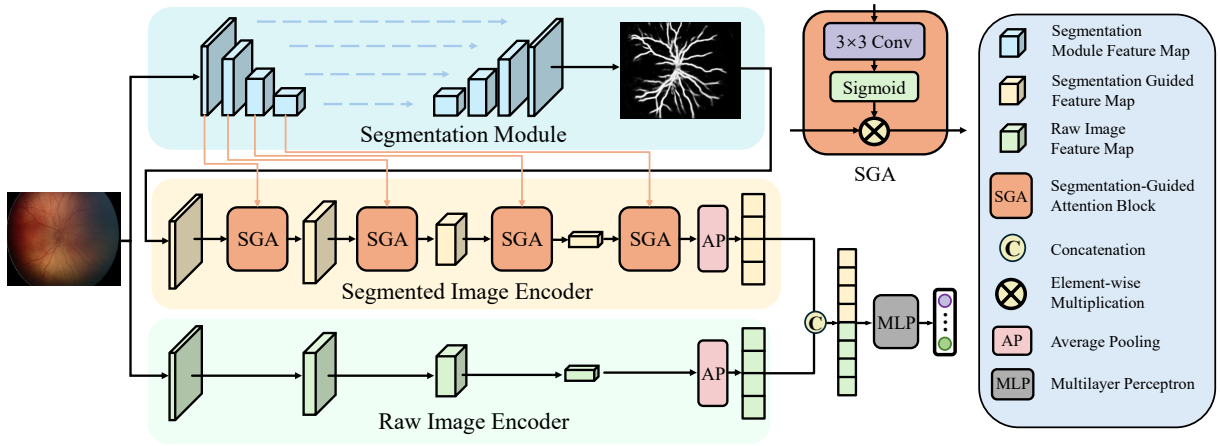

Figure 2: Overview of SegImgNet. The input retinal image is first processed by the segmentation module to generate segmentation maps (top). These segmentation maps, along with the raw retinal image, are then fed into separate encoders to extract disease-related features (bottom). The SGA block leverages intermediate feature maps from the segmentation module to generate attention maps, guiding the segmented image encoder's focus on retinal structural features. Finally, the classifier integrates outputs from both encoders for disease classification.

sequently applies multiple machine learning algorithms to the segmented data for disease prediction. Although these methods improve attention to segmented regions, they ignore valuable information from complementary image areas, potentially limiting overall diagnostic performance.

To extract more comprehensive features, recent studies have integrated both segmentation results and retinal images for disease diagnoses (Alam et al. 2023; Joshi, Sharma, and Dutta 2024; Xiong et al. 2025). Specifically, some approaches fuse segmented and raw images into a single input and then feed it into an encoder for classification, as shown in Figure 1(c), while others process segmented and raw images through separate encoders to extract features for disease classification, as shown in Figure 1(d). For example, (Alam et al. 2023) stacks segmentation maps and retinal images into a single input for GoogleNet, whereas VisionDeep-AI (Joshi, Sharma, and Dutta 2024; Xiong et al. 2025) concatenates features extracted from segmented images and retinal images using separate EfficientNet or ResNet50 models. However, these methods lack explicit interactions between segmentation and classification feature spaces. As a result, retinal anatomical features are not fully leveraged to enhance the learned representations in the classification model, limiting the model's ability to incorporate prior structural information for improved disease diagnosis.

In this paper, we propose SegImgNet, a deep learning framework for retinal disease classification that integrates both retinal images and segmentation maps. By leveraging multi-scale structural feature maps obtained from segmentation along with original retinal images, SegImgNet enhances classification performance. The framework consists of two main components: a segmentation module and a classification module. The segmentation module, based on the U-Net (Ronneberger, Fischer, and Brox 2015) architecture, generates retinal structure feature maps. The classification module includes a segmented image encoder, a raw image encoder, a classifier, and Segmentation-Guided Attention (SGA) blocks. The segmented image encoder extracts disease-related local features, while the raw image encoder captures broader global contextual information. Both encoders are built on the ConvNeXt architecture, and the classifier combines their outputs into a unified representation for disease classification. Additionally, the SGA block enhances feature extraction by generating attention maps from structural segmentation, allowing the model to focus on critical retinal details. Extensive experiments on public AIROGS and private e-ROP datasets demonstrate that SegImgNet consistently outperforms existing state-of-the-art methods for retinal disease diagnosis.

Our Approach

Figure 2 illustrates the architecture of SegImgNet, which consists of two main components: a segmentation module and a classification module. The details of these two modules are introduced below.

Segmentation Module

The segmentation module $f_{seg}(\cdot)$ employs a U-Net architecture to generate retinal structure feature maps. U-Net utilizes a symmetric encoder-decoder architecture with skip connections, enabling it to capture both low-level spatial details and high-level abstract features. This structure ensures the precise localization of the retinal structures while preserving fine-grained anatomical details.

The U-Net's encoder consists of multiple convolutional layers followed by downsampling operations, progressively reducing spatial resolution while enhancing feature abstraction. This hierarchical representation enables the model to capture retinal structures across multiple scales, which is essential for detecting both fine-grained details and broader pathological patterns. The decoder, on the other hand, reconstructs the segmented image by gradually upsampling

the encoded features, restoring spatial details lost during downsampling. Skip connections bridge the corresponding encoder and decoder layers, allowing high-resolution features from the encoder to be directly merged with upsampled features in the decoder. These connections help preserve fine-grained structural information, which is crucial for accurately delineating retinal regions.

Specifically, given a retinal image $\mathbf{x} \in \mathbb{R}^{H\times W\times C_{raw}}$, where H, W, and C_{raw} denote the height, width, and channel size of the raw image, respectively, the corresponding segmented image $\mathbf{x}_{seg} \in \mathbb{R}^{H\times W\times C_{seg}}$ and multi-scale retinal structural feature maps $\{\mathbf{h}_{seg}^{(i)}\}_{i=1}^{L} \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times C_i}$ are obtained as follows:

$$\mathbf{x}_{seg}, \{\mathbf{h}_{seg}^{(i)}\}_{i=1}^{L} = f_{seg}(\mathbf{x}), \tag{1}$$

where C_{seg} represents the channel size of the segmented image, and L represents the number of feature scales, which is empirically set to 4 in this study (Li et al. 2024).

Classification Module

The classification module extracts structural features from segmentation maps and contextual representations from raw retinal images for disease diagnoses. It consists of a segmented image encoder, a raw image encoder, a classifier, and SGA blocks. Each component is detailed below.

Segmented Image Encoder: The segmented image encoder extracts fine-grained structural representations from the output of the segmentation module while incorporating segmentation priors at multiple stages. Here we use ConvNeXt (Liu et al. 2022) as a feature extractor or backbone for this encoder. Each stage of the feature extractor is equipped with a Segmentation-Guided Attention (SGA) block, which enhances attention to retinal structural features. By selectively emphasizing relevant features and filtering out less informative regions, the SGA block ensures that the extracted representations retain critical anatomical details essential for accurate disease classification and improved diagnostic reliability. The final segmentation map feature representations are obtained from the last stage of the feature extractor, where segmentation-guided information is further enriched with anatomical details.

Specifically, the SGA block builds on the approach in (Li et al. 2024), utilizing convolution operations and a sigmoid activation function to refine feature extraction. It enhances the intermediate feature maps of the segmented image encoder by integrating segmentation-derived structural information. Given the output feature map $\mathbf{h}_{local}^{(i)}$ from the i-th stage of the feature extractor and the corresponding retinal structural feature map $\mathbf{h}_{seg}^{(i)}$, the SGA block produces an enhanced representation, formulated as:

$$\tilde{\mathbf{h}}_{local}^{(i)} = \sigma(\text{Conv}_{3\times3}(\mathbf{h}_{seg}^{(i)})) \odot \mathbf{h}_{local}^{(i)} \tag{2}$$

where $\text{Conv}_{3\times3}(\cdot)$ represents a convolutional layer with a kernel size of 3×3 used for adjusting $\mathbf{h}_{seg}^{(i)}$ spatial dimensions to match $\mathbf{h}_{local}^{(i)}$ size, $\sigma(\cdot)$ denotes the sigmoid activation function to generate the attention score, and $\odot$ denotes element-wise product to highlight the retinal structure part of feature map. The resulting enhanced feature map $\tilde{\mathbf{f}}_{local}^{(i)}$ is then fed into the next stage of the feature extractor.

Raw Image Encoder: The raw image encoder is designed to extract global contextual representations from retinal images, complementing the structural features extracted by the segmented image encoder. Similarly to the segmented image encoder, it employs ConvNeXt as the backbone. However, unlike the segmented image encoder, which processes segmented images with segmentation-derived feature map enhancement, the raw image encoder focuses on capturing broader disease-relevant patterns within the retinal image. In particular, the raw image encoder is not equipped with SGA blocks, ensuring that it does not emphasize the same structural features as the segmented image encoder. This design preserves feature complementarity by allowing the segmented image encoder to prioritize segmentation-guided structural information. The final global feature representations are obtained from the deepest stage of ConvNeXt, where high-level disease-relevant information is encoded while retaining spatial context.

Classifier: After obtaining the segmented image feature embedding $\mathbf{h}_{local}$ and the raw image feature embedding $\mathbf{h}_{global}$ from the encoders, the classifier concatenates them to form a comprehensive feature embedding $\mathbf{h}_{cls}$ for disease classification. It then applies a Multilayer Perceptron (MLP) followed by a $softmax$ activation function to classify diseases based on the feature embedding $\mathbf{h}_{cls}$. Specifically, the probability of the k-th disease, $\hat{y}_k$, is computed as follows:

$$\hat{y}_k = \frac{\exp(f_{MLP}^{k}(\mathbf{h}_{cls}))}{\sum_{j=1}^{K}\exp(f_{MLP}^{j}(\mathbf{h}_{cls}))}, \tag{3}$$

where K denotes the total number of classes, and $f_k(\mathbf{h}_{cls}))$ represents the MLP output for class k.

Overall Loss Function

To address the class imbalance commonly found in medical datasets, we employ a Weighted Cross-Entropy (WCE) loss function to train SegImgNet. This loss function assigns higher penalties to misclassified minority-class samples, mitigating the dominance of majority classes and improving the model's ability to detect rare disease cases. The WCE loss is defined as:

$$\mathcal{L}_{\text{WCE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} w_k \cdot y_k^{(i)} \log \hat{y}_k^{(i)} \quad \text{s.t.} \sum_{k=1}^{K} w_k = 1, \tag{4}$$

where N is the number of input samples, w_k denotes the weight assigned to class k. $y_k^{(i)}$ and $\hat{y}_k^{(i)}$ represent the one-hot encoded ground truth label and the predicted probability of the sample i, respectively.

Experiment

Experimental Setup

Datasets: We evaluated SegImgNet on two datasets: the public AIROGS dataset and the private e-ROP dataset.

Dataset	Method	RAW	SEG	AUC	Sensitivity	Specificity	F1 score	Precision	Accuracy
AIROGS	ResNet50	✓	✗	0.969±0.006	0.914±0.012	0.921±0.015	0.917±0.008	0.920±0.013	0.917±0.008
	RETFound	✓	✗	0.925±0.004	0.807±0.032	0.879±0.031	0.837±0.008	0.872±0.025	0.843±0.005
	ResNet50-MaxViT	✓	✗	0.970±0.004	0.925±0.013	0.907±0.018	0.917±0.006	0.909±0.015	0.916±0.007
	AVS-DenseNet	✗	✓	0.850±0.006	0.859±0.013	0.698±0.009	0.795±0.010	0.740±0.008	0.778±0.010
	Res-Unet-CNNs	✗	✓	0.890±0.006	0.797±0.037	0.824±0.032	0.807±0.013	0.820±0.021	0.811±0.010
	U-Nets-DenseNet	✗	✓	0.915±0.004	0.823±0.036	0.846±0.032	0.832±0.010	0.844±0.022	0.834±0.006
	SA-GoogleNet	✓	✓	0.961±0.003	0.900±0.030	0.901±0.027	0.900±0.007	0.902±0.022	0.900±0.006
	Multi-GlaucNet	✓	✓	0.965±0.003	0.949±0.014	0.844±0.032	0.902±0.009	0.860±0.023	0.897±0.010
	VisionDeep-AI	✓	✓	0.970±0.002	0.930±0.012	0.899±0.004	0.916±0.005	0.902±0.003	0.914±0.005
	SegImgNet	✓	✓	**0.985±0.001**	**0.949±0.010**	**0.934±0.009**	**0.941±0.002**	**0.935±0.008**	**0.940±0.003**
e-ROP	ResNet50	✓	✗	0.895±0.007	0.782±0.055	0.842±0.028	0.546±0.019	0.423±0.033	0.834±0.019
	RETFound	✓	✗	0.854±0.017	0.725±0.024	0.827±0.012	0.497±0.018	0.378±0.018	0.814±0.011
	ResNet50-MaxViT	✓	✗	0.901±0.010	0.812±0.040	0.825±0.026	0.540±0.020	0.406±0.028	0.823±0.019
	AVS-DenseNet	✗	✓	0.805±0.021	0.780±0.032	0.677±0.031	0.391±0.025	0.261±0.020	0.690±0.028
	Res-Unet-CNNs	✗	✓	0.831±0.009	0.706±0.030	0.796±0.015	0.449±0.004	0.332±0.007	0.784±0.009
	U-Nets-DenseNet	✗	✓	0.883±0.012	0.732±0.038	0.868±0.017	0.555±0.021	0.449±0.028	0.851±0.012
	SA-GoogleNet	✓	✓	0.827±0.027	0.757±0.044	0.736±0.067	0.431±0.049	0.305±0.055	0.739±0.055
	Multi-GlaucNet	✓	✓	0.867±0.021	0.771±0.070	0.803±0.044	0.496±0.021	0.368±0.030	0.799±0.030
	VisionDeep-AI	✓	✓	0.893±0.007	0.731±0.028	**0.885±0.010**	0.540±0.021	0.454±0.023	**0.865±0.010**
	SegImgNet	✓	✓	**0.921±0.006**	**0.831±0.027**	0.843±0.042	**0.589±0.015**	**0.465±0.025**	0.857±0.013

Table 1: Performance comparison of classification models (mean ± std) using raw retinal images (RAW), segmented images (SEG), or both on AIROGS and e-ROP datasets, with bold indicating the best performance.

Dataset	Model Configurations	AUC	Sensitivity	Specificity	F1 score	Precision	Accuracy
AIROGS	w/o segmented image encoder	0.984±0.002	0.947±0.015	0.929±0.019	0.939±0.003	0.931±0.017	0.938±0.006
	w/o raw image encoder	0.955±0.003	0.863±0.025	0.903±0.026	0.880±0.009	0.900±0.021	0.883±0.008
	w/o SGA	0.984±0.002	0.942±0.009	0.930±0.011	0.937±0.003	0.932±0.009	0.937±0.004
	Full SegImgNet	**0.985±0.001**	**0.949±0.010**	**0.934±0.009**	**0.941±0.002**	**0.935±0.008**	**0.940±0.003**
e-ROP	w/o segmented image encoder	0.908±0.010	0.810±0.029	0.841±0.035	0.561±0.034	0.432±0.050	0.837±0.027
	w/o raw image encoder	0.892±0.019	0.793±0.069	0.827±0.038	0.533±0.028	0.405±0.040	0.823±0.026
	w/o SGA	0.914±0.005	0.825±0.046	0.831±0.034	0.555±0.024	0.422±0.044	0.831±0.024
	Full SegImgNet	**0.921±0.006**	**0.831±0.027**	**0.843±0.042**	**0.589±0.015**	**0.465±0.025**	**0.857±0.013**

Table 2: Ablation study of SegImgNet components on AIROGS and e-ROP datasets (mean ± std).

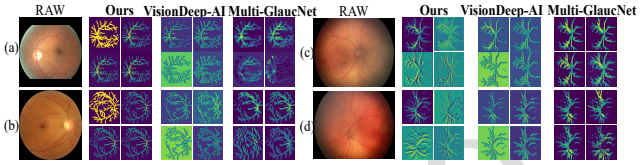

Figure 3: Intermediate feature map visualizations of top-3 methods. (a) and (b) are from the AIROGS dataset, where (a) is healthy and (b) is glaucomatous. (c) and (d) are from the e-ROP dataset, where (c) is healthy and (d) is ROP.

- The AIROGS dataset (Steen et al. 2023) is an improved glaucoma dataset consisting of a balanced subset of standardized retinal images. It is derived from the Rotterdam EyePACS AIROGS set, which contains 113,893 color retinal images from 60,357 subjects across approximately 500 different sites with heterogeneous ethnicities. These retinal images were labeled as glaucomatous or healthy based on clinical evaluations performed by glaucoma specialists. For this study, we used 4,950 publicly available retinal images, including 2,475 glaucomatous images and 2,475 healthy images.
- The e-ROP dataset originates from the Telemedicine Methods for Evaluating Acute Retinopathy of Prematurity (e-ROP) study (Quinn et al. 2014), which collected retinal images from 1,257 infants admitted to neonatal intensive care units across 13 centers in North America. These images are captured using wide-angle retinal cameras during scheduled diagnostic examinations. Each retinal image was labeled as either preandplus or normal by experienced ophthalmologists. In this study, we used 7,811 center-view retinal images, including 990 preandplus images and 6,821 normal images.

Baselines: We evaluated our proposed model against a diverse set of state-of-the-art classification models, categorized based on the type of input used for disease classification: (1) Retinal image-based models, which classify diseases using only raw retinal images, including ResNet50 (Huang et al. 2023), RETFound (Zhou et al. 2023), and ResNet50-MaxViT (Zhao et al. 2023); (2) Segmented image-based models, which rely only on segmented images, such as AVS-DenseNet (Almeida et al. 2024), Res-Unet-CNNs (Wang et al. 2021a), and U-Nets-DenseNet (Li and Liu 2022); and (3) Hybrid models, which integrate both retinal images and segmented images, including SA-GoogleNet (Alam et al. 2023), Multi-GlaucNet (Xiong et al. 2025), and VisionDeep-AI (Joshi, Sharma, and Dutta 2024).

Evaluation Metrics: We evaluated model performance us-

ing six standard metrics: Area Under the Receiver Operating Characteristic Curve (AUC) to assess discriminative ability, sensitivity (true positive rate) to quantify disease detection capability, specificity (true negative rate) to measure the ability to identify healthy cases, precision (positive predictive value) to evaluate diagnostic confidence, F1-score to balance precision and recall, and accuracy to reflect overall classification performance.

Implementation Details: To ensure a fair comparison, we conducted five-fold cross-validation on each dataset, partitioning the labeled images into 80% training data and 20% test data. The training data was further divided into a training set and a validation set in a ratio $3:1$, maintaining the original class distribution for hyperparameter tuning. To mitigate class imbalance in the training set, we employed the Random OverSampling Examples (ROSE) (Hayaty, Muthmainah, and Ghufran 2020) technique to balance the number of images in each class. Additionally, we applied data-augmentation techniques, including image flipping, cropping, and scaling, to the training set to improve the model's generalization ability. For consistency, all retinal images were resized to 256×256 pixels.

All compared models are implemented using the PyTorch framework. The segmentation components were pre-trained on 933 samples from six public retinal vessel segmentation datasets: FIVES (Jin et al. 2022), DRIVE (Staal et al. 2004), STARE (Hoover, Kouznetsova, and Goldbaum 2000), CHASEDB1 (Budai et al. 2013a), HRF (Budai et al. 2013b), and Retinal Blood Vessel Segmentation (Wang et al. 2021b). The classification components were pre-trained on the ImageNet dataset, except for RETFound, which was trained on its custom dataset.

All experiments were accelerated using NVIDIA RTX A5000 GPUs. Model optimization was performed using the Adam optimizer. To enhance performance, we conducted a grid search to fine-tune key hyperparameters, including the learning rate, batch size, and disease class weight in the weighted cross-entropy loss function. The learning rate was explored within the range 5×10^{-5} to 1×10^{-3}, batch sizes were selected from $\{16, 32, 64, 128\}$, and class weight were varied from 0.5 to 0.9 with a step size of 0.1. We set the maximum number of training epochs to 200, with early stopping applied if validation performance did not improve within 20 epochs. The best-performing model checkpoint on the validation set was selected for testing.

Experimental Results

Comparisons with Baselines: Table 1 presents the disease classification performance of all compared models across two datasets. Specifically, we have the following observations: SegImgNet consistently outperforms all baselines across key metrics on both datasets, demonstrating its superior capability to distinguish between disease and normal cases. While SegImgNet achieves slightly lower specificity (0.843 ± 0.042) and accuracy (0.857 ± 0.013) compared to VisionDeep-AI (0.885 ± 0.010 and 0.865 ± 0.010, respectively) on the e-ROP dataset, it remains highly competitive on these two metrics. More importantly, while VisionDeep-AI exhibits higher specificity and accuracy, it falls short in other critical metrics, particularly sensitivity (0.731 ± 0.028 for VisionDeep-AI vs. 0.831 ± 0.027 for SegImgNet). This lower sensitivity increases the risk of missed diagnoses, which can lead to delayed treatment. In medical applications, sensitivity is crucial, as missing a disease diagnosis can have far more severe consequences than misclassifying a healthy individual. Notably, SegImgNet achieves the highest sensitivity among all baselines on both data sets, confirming its effectiveness in clinical decision-making. Figure 3 shows the visualization of intermediate feature maps from the segmented image encoder of the top three models (SegImgNet, VisionDeep-AI and Multi-GlaucNet) across two datasets. We selected the feature maps produced by each model's second downsampling layer and visualized four representative channels, chosen based on their mean and variance. The visualization results demonstrate that our approach achieves higher structural clarity and consistency compared to the other two approaches. Specifically, SegImgNet more distinctly delineates prominent edges and anatomical structures, thereby enhancing its capability to preserve and highlight morphological features for accurate retinal analysis.

Ablation Study: Here we investigated the contribution of each key component in SegImgNet, including the segmented image encoder, raw image encoder, and SGA block. Table 2 presents the performance of different model variants: "w/o segmented image encoder" excludes the segmented image encoder, "w/o raw image encoder" removes the raw image encoder, and "w/o SGA" omits the SGA block. The results demonstrate that each component is essential for optimal performance. Removing the segmented image encoder significantly reduces the model's ability to capture retinal structural features, while eliminating the raw image encoder weakens its capacity to extract global contextual information. Furthermore, the absence of the SGA block degrades classification performance, highlighting the importance of multi-scale retinal structural feature maps in enhancing representation learning. The complete SegImgNet model, incorporating all components, achieves the highest performance, emphasizing the importance of integrating local and global feature extraction with attention-based enhancement. These findings confirm that each module plays a critical role in maximizing disease classification accuracy.

Conclusion

In this study, we introduce SegImgNet, a deep learning model that integrates local retinal structural features from segmented images with global contextual information from raw images for disease classification. Extensive experiments on public and private datasets show that SegImgNet outperforms existing methods, demonstrating the effectiveness of segmentation-guided attention for feature enhancement. Our findings highlight the potential of incorporating retinal structural priors into deep learning frameworks to improve the robustness of AI-driven medical imaging. Future work will focus on optimizing feature fusion, expanding the model to broader ophthalmic applications, and improving generalization across diverse clinical datasets.

Acknowledgments

This study was in part supported by the National Institutes of Health (R21EY034179 and P30-EY01583-26) and Research to Prevent Blindness Foundation (Research to Prevent Blindness (RPB), as well as the support of NSF MRI grant (2215789) for our computing infrastructure.

References

Alam, M.; Zhao, E. J.; Lam, C. K.; and Rubin, D. L. 2023. Segmentation-assisted fully convolutional neural network enhances deep learning performance to identify proliferative diabetic retinopathy. *Journal of Clinical Medicine*, 12(1): 385.

Almeida, J.; Kubicek, J.; Penhaker, M.; Cerny, M.; Augustynek, M.; Varysova, A.; Bansal, A.; and Timkovic, J. 2024. Enhancing ROP Plus Form Diagnosis: An Automatic Blood Vessel Segmentation Approach for Newborn Fundus Images. *Results in Engineering*, 103054.

Budai, A.; Bock, R.; Maier, A.; Hornegger, J.; and Michelson, G. 2013a. Robust vessel segmentation in fundus images. *International journal of biomedical imaging*, 2013(1): 154860.

Budai, A.; Bock, R.; Maier, A.; Hornegger, J.; and Michelson, G. 2013b. Robust vessel segmentation in fundus images. *International journal of biomedical imaging*, 2013(1): 154860.

Hayaty, M.; Muthmainah, S.; and Ghufran, S. M. 2020. Random and synthetic over-sampling approach to resolve data imbalance in classification. *International Journal of Artificial Intelligence Research*, 4(2): 86–94.

Hoover, A.; Kouznetsova, V.; and Goldbaum, M. 2000. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical imaging*, 19(3): 203–210.

Huang, Y.; Lin, L.; Cheng, P.; Lyu, J.; Tam, R.; and Tang, X. 2023. Identifying the key components in resnet-50 for diabetic retinopathy grading from fundus images: a systematic investigation. *Diagnostics*, 13(10): 1664.

Jin, K.; et al. 2022. Fives: A fundus image dataset for artificial Intelligence based vessel segmentation, Figshare.

Joshi, R. C.; Sharma, A. K.; and Dutta, M. K. 2024. VisionDeep-AI: Deep learning-based retinal blood vessels segmentation and multi-class classification framework for eye diagnosis. *Biomedical Signal Processing and Control*, 94: 106273.

Li, C.; Wang, R.; He, P.; Chen, W.; Wu, W.; and Wu, Y. 2024. Segmentation prompts classification: A nnUNet-based 3D transfer learning framework with ROI tokenization and cross-task attention for esophageal cancer T-stage diagnosis. *Expert Systems with Applications*, 258: 125067.

Li, H.; Cao, J.; Grzybowski, A.; Jin, K.; Lou, L.; and Ye, J. 2023. Diagnosing systemic disorders with AI algorithms based on ocular images. In *Healthcare*, volume 11, 1739.

Li, P.; and Liu, J. 2022. Early diagnosis and quantitative analysis of stages in retinopathy of prematurity based on deep convolutional neural networks. *Translational Vision Science & Technology*, 11(5): 17–17.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986.

Quinn, G. E.; Ying, G.-s.; Daniel, E.; Hildebrand, P. L.; Ells, A.; Baumritter, A.; Kemper, A. R.; Schron, E. B.; Wade, K.; e ROP Cooperative Group; et al. 2014. Validity of a telemedicine system for the evaluation of acute-phase retinopathy of prematurity. *JAMA ophthalmology*, 132(10): 1178–1184.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Sivapriya, G.; Devi, R. M.; Keerthika, P.; and Praveen, V. 2024. Automated diagnostic classification of diabetic retinopathy with microvascular structure of fundus images using deep learning method. *Biomedical Signal Processing and Control*, 88: 105616.

Staal, J.; Abràmoff, M. D.; Niemeijer, M.; Viergever, M. A.; and Van Ginneken, B. 2004. Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging*, 23(4): 501–509.

Steen, J.; Kiefer, R.; Ardali, M.; Abid, M.; and Amjadian, E. 2023. Standardized and Open-Access Glaucoma Dataset for Artificial Intelligence Applications. *Investigative Ophthalmology & Visual Science*, 64(8): 384–384.

Tan, Y. Y.; Kang, H. G.; Lee, C. J.; Kim, S. S.; Park, S.; Thakur, S.; Da Soh, Z.; Cho, Y.; Peng, Q.; Lee, K.; et al. 2024. Prognostic potentials of AI in ophthalmology: systemic disease forecasting via retinal imaging. *Eye and Vision*, 11(1): 17.

Wang, J.; Ji, J.; Zhang, M.; Lin, J.-W.; Zhang, G.; Gong, W.; Cen, L.-P.; Lu, Y.; Huang, X.; Huang, D.; et al. 2021a. Automated explainable multidimensional deep learning platform of retinal images for retinopathy of prematurity screening. *JAMA network open*, 4(5): e218758–e218758.

Wang, J.; Ji, J.; Zhang, M.; Lin, J.-W.; Zhang, G.; Gong, W.; Cen, L.-P.; Lu, Y.; Huang, X.; Huang, D.; et al. 2021b. Automated explainable multidimensional deep learning platform of retinal images for retinopathy of prematurity screening. *JAMA network open*, 4(5): e218758–e218758.

Xiong, H.; Long, F.; Alam, M. S.; and Sang, J. 2025. Multi-GlaucNet: A multi-task model for optic disc segmentation, blood vessel segmentation and glaucoma detection. *Biomedical Signal Processing and Control*, 99: 106850.

Zhao, J.; Lei, H.; Xie, H.; Li, P.; Liu, Y.; Zhang, G.; and Lei, B. 2023. Dual-Branch Attention Network and Swin Spatial Pyramid Pooling for Retinopathy of Prematurity Classification. In *ISBI*, 1–4.

Zhou, Y.; Chia, M. A.; Wagner, S. K.; Ayhan, M. S.; Williamson, D. J.; Struyven, R. R.; Liu, T.; Xu, M.; Lozano, M. G.; Woodward-Court, P.; et al. 2023. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981): 156–163.