Enhancing Fundus Image Classification with Semantic Segmentation-Based Attention Mask

Elena Himbitskaya
Belarusian State University
Belarusian State Medical University
Minsk, Belarus
Email: fpm.gimbicka@bsu.by

Kseniya Svistunova
Belarusian State University
Minsk, Belarus
Email: fpm.svistuno@bsu.by

Grigory Karapetyan
Belarusian State Medical University
Minsk, Belarus
Email: greg.itlab@gmail.com

Alexander Nedzved
Belarusian State University
United Institute of Informatics Problems NAS Belarus
Minsk, Belarus
Email: Anedzved@bsu.by

Sergey Ablameyko
Belarusian State University
National Academy of Sciences of Belarus
Minsk, Belarus
Email: ablameyko@bsu.by

***Abstract*—This research proposes a method for classifying ocular diseases in fundus images using semantic segmentation as an attention mechanism. Unlike conventional approaches that rely on the entire retinal image, the proposed framework emphasizes anatomically relevant regions extracted via segmentation of the optic disc, optic cup, and retinal vessels. These segmentation masks are integrated into the classification pipeline to enhance feature learning. A EfficientNetB6-based classifier is utilized to evaluate the impact of this strategy. Experimental results demonstrate improvements in classification performance across multiple evaluation metrics.**

***Keywords*—automated model, ocular fundus, deep learning, attention mask, semantic segmentation, optic disc, optic cup, vessels, classification, UNet, EfficientNetB6**

I. Introduction

Fundus imaging has become a cornerstone in ophthalmology, offering non-invasive and high-resolution visualization of the retina and its associated structures. These images are integral in diagnosing a broad spectrum of ocular and systemic diseases. The interpretation of such images, however, demands a high level of expertise and can be time-consuming, especially in regions with limited access to ophthalmic specialists [1]. As the global burden of ocular diseases continues to rise, there is a critical need for automated systems that can support clinical decision-making and screening efforts.

Recent advancements in deep learning have shown considerable promise in addressing this challenge [2]. Convolutional Neural Networks (CNNs) have demonstrated strong performance in a variety of medical image classification tasks, including retinal disease recognition [3]. Nevertheless, one of the limitations of end-to-end classification models is their reliance on global image features, which may dilute the impact of localized, clinically relevant structures. Consequently, classification performance may suffer, particularly in the early stages of disease when only subtle signs are present.

To improve the focus of deep learning models on diagnostically relevant regions, attention mechanisms and region-guided methods have gained popularity. In this study, we explore the use of semantic segmentation as an attention strategy in the classification of ocular fundus images. By segmenting anatomical regions known to exhibit pathological changes—such as the optic disc, cup, and retinal vessels—and using these masks to enhance or filter the input images, we aim to guide the classifier's attention to the most informative areas [4].

II. Sympoms of ocular diseases in fundus imaging

Fundus photography enables non-invasive visualization of the internal structures of the eye, including the retina, optic disc, macula, and posterior pole. It plays a critical role in the identification and monitoring of numerous ocular diseases. The following section outlines key pathological signs of common ocular diseases as they appear in retinal fundus images.

A. Myopia

Pathological myopia is typically associated with axial elongation of the eyeball, which leads to mechanical stretching and thinning of the retina. Fundus images of myopic patients frequently exhibit peripapillary atrophy, tessellated fundus appearance, tilted optic discs, and staphylomas. These features reflect structural deformation and progressive degeneration of the posterior segment, which can predispose the eye to chorioretinal atrophy and retinal detachment.

B. Hypertension

Hypertensive retinopathy results from chronic elevated blood pressure and manifests through various microvas-

cular changes. Common fundus signs include generalized and focal arteriolar narrowing, arteriovenous (AV) nicking, arteriolar wall opacification (copper wiring and silver wiring), flame-shaped hemorrhages, cotton wool spots, and hard exudates. In severe cases, swelling of the optic disc and macular star formation may occur, indicating malignant hypertension and necessitating immediate intervention.

C. Diabetes

Diabetic retinopathy (DR) is one of the most prevalent causes of blindness globally. Its hallmark features in fundus images include microaneurysms, intraretinal hemorrhages (dot and blot types), hard exudates, cotton wool spots, and retinal edema. Proliferative diabetic retinopathy (PDR) may present with neovascularization on the optic disc or elsewhere in the retina, vitreous hemorrhage, and tractional retinal detachment. Diabetic macular edema, characterized by retinal thickening in the macular area, is a leading cause of vision loss in DR [5].

D. Glaucoma

Glaucoma is characterized by progressive optic neuropathy and loss of retinal ganglion cells, with corresponding changes visible in fundus photographs. These include increased cup-to-disc (C/D) ratio, thinning or notching of the neuroretinal rim, peripapillary atrophy, and optic disc hemorrhages (especially in normal-tension glaucoma). Advanced stages may show "bean-pot" excavation of the optic nerve head. Evaluation of the C/D ratio and asymmetry between eyes is essential in glaucoma detection.

E. Cataract

Although cataract primarily affects the crystalline lens and is best visualized via slit-lamp biomicroscopy, it can have indirect effects on fundus photography. Opacification of the lens leads to decreased image contrast, blurring, and reduced visibility of retinal structures. In fundus images, this appears as a generalized haze, particularly in the red channel, which may complicate retinal assessment and affect automated analysis accuracy.

F. Age-Related Macular Degeneration (AMD)

AMD is a degenerative disease of the central retina and is classified into dry (non-exudative) and wet (exudative) forms. Early signs in fundus images include drusen (yellowish extracellular deposits beneath the retina), pigmentary changes, and geographic atrophy. In neovascular AMD, subretinal hemorrhage, fluid accumulation, and choroidal neovascular membranes may be observed. These manifestations often result in central vision loss and are identifiable through high-resolution fundus imaging.

The specific appearance of these pathologies on fundus images forms the basis for automated diagnostic algorithms. Accurate segmentation and localization of relevant anatomical structures will later enable our models to focus on informative regions.

III. Multiclass semantic segmentation methods

To extract the most informative features we segmented the main objects:

- optic disc;
- optic cup;
- blood vessels.

To form a semantic map of the image's most informative objects was chosen a convolutional neural network of Unet widely used for segmentation of medical images.

A. Optic disc and optic cup segmentation

The segmentation model employed for segmenting optic disc and optic cup is based on a modified U-Net architecture implemented in PyTorch. The network is trained to produce dual-channel output masks corresponding to the optic disc and optic cup regions.

The model follows the classical U-Net design with symmetric encoder-decoder paths and skip connections between corresponding layers. Each encoder block performs two convolutions with ReLU activation, followed by a downsampling operation (MaxPooling). The decoder mirrors this process with upsampling via transposed convolutions and concatenation with features from the encoder.

- Input: RGB image of shape (3, 256, 256)
- Output: Segmentation mask of shape (2, 256, 256) (channel 0: disc, channel 1: cup)

Table I
Semantic interpretation of multi-channel segmentation mask

Channel 0 (Disc)	Channel 1 (Cup)	Semantic Class
0	0	Background
1	0	Disc only
0	1	Cup only
1	1	Cup inside Disc

The training was supervised using a Binary Cross-Entropy (BCE) loss per class for 30 epochs with early stopping. The loss during training epochs can be seen on Figure 1.

Model performance was evaluated during training using the Dice coefficient (Figure 5), which provides a measure of spatial overlap between predicted and true segmentation masks.

The final evaluation on the validation set showed a high quality of segmentation. The Binary IoU (Jaccard index) achieved:

- **Optic disc**: 0.92
- **Optic cup**: 0.83

These values indicate good spatial agreement between predictions and ground truth masks, especially for the

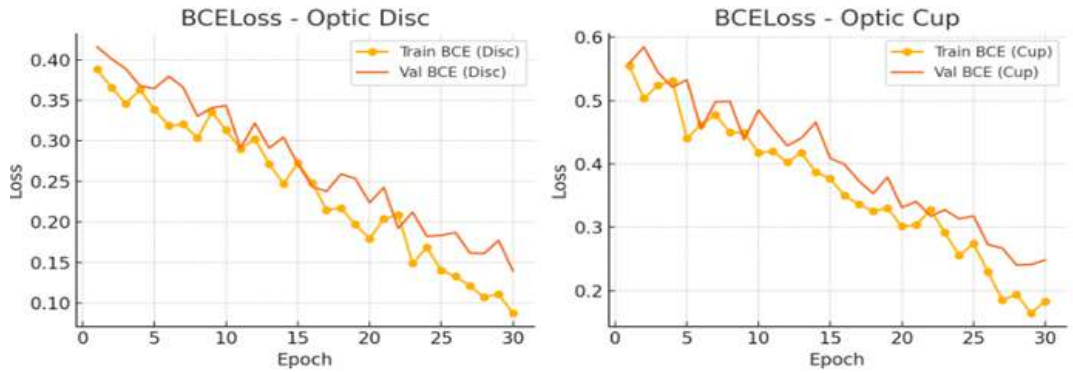

Figure 1. BCELoss during training for optic disc and optic cup.

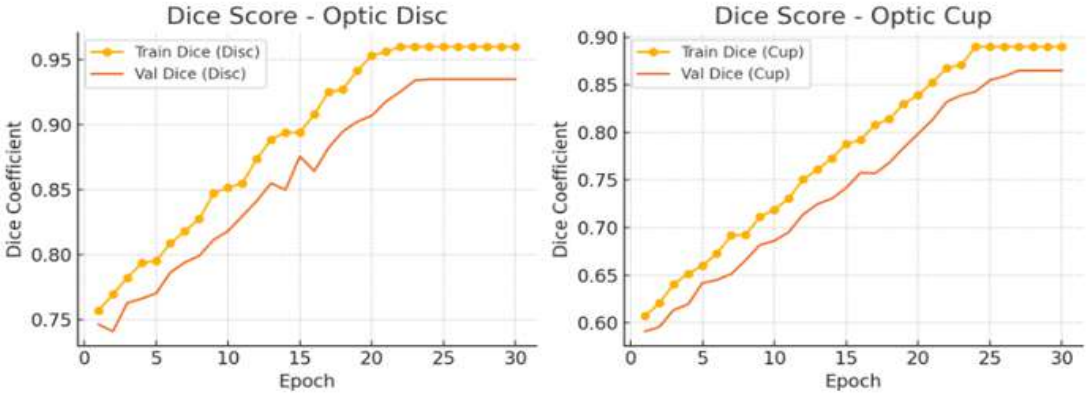

Figure 2. Dice coefficient during training for optic disc and optic cup.

optic disc. The optic cup, being smaller and less distinct in fundus images, showed a lower but still acceptable IoU score.

The Dice coefficient at convergence was approximately 0.95 for the optic disc and 0.88 for the optic cup, which reflects the class imbalance and visual complexity in segmenting the excavation zone.

An analysis of training dynamics shows that the loss curves are not monotonic and include fluctuations, particularly for the optic cup, which suggests sensitivity to anatomical variations and class imbalance.

Further improvement of segmentation accuracy can be achieved by:

- Incorporating *Dice loss* or *combined BCE + Dice loss* during training to better handle imbalanced regions;
- Using *focal loss* to reduce the impact of background pixels;
- Applying *data augmentation* focused on enhancing variability in cup morphology.

Overall, the segmentation model is robust and provides sufficiently accurate anatomical masks for our algorithm.

B. Segmentation of blood vessels

The training of the model extracting the vascular network of the image was carried out in 2 stages. In the first stage, the network was trained on an additional set of 300 labeled data from publicly available datasets such as DRIVE [6], CHASE DB1 [7] and HRF [8]. In the second stage, the network was trained on target images.

Initially, the analyzed three-channel (RGB) image was compressed to a size of 996 x 996. After that, it was split into 9 slices with a resolution of 352 x 352 so that each slice captures a part of the neighboring slices (10 pixels). This is to eliminate distortion at the boundary between two tiles. Vessel segmentation by the neural network was performed for each tile. We chose UNet architecture with a resnet18 backbone pre-trained on the ImageNet dataset.

The model contains 23 convolutional layers and consists of convolutional (encoder) and up-convolutional (decoder) parts. To reduce each 64-component vector to the required number of classes, 1×1 convolutions are applied on the last layer. The input image size is determined by the need for even values of height and width for adequate application of subsampling operation (2×2 max pooling).

The network is trained by stochastic gradient descent based on the input images and their corresponding segmentation maps (masks). Applied function, soft-max brings the model prediction to the mask view. The loss

function is a binary cross-entropy + jaccard functions. The accuracy is calculated by the BinaryIOU() [9] function, which finds the ratio of the correctly predicted mask to the union of the predicted and true masks. After the tiles were merged into a single image with the boundary 5 pixels cropped on each of them. They were then merged into rows. To smooth the transition between two tiles, their 5-pixel boundaries are overlaid and the resulting brightness is calculated using alpha blending to obtain a smooth transition. This process is shown in Fig. 3.

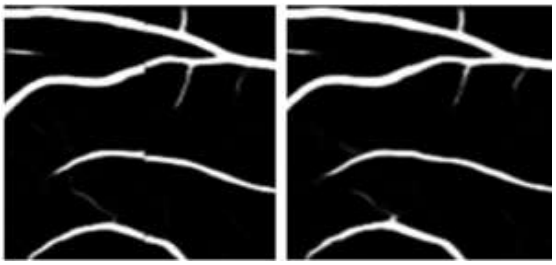

Figure 3. left – result of segmentation model on neighboring tiles, right – tiles merged with alpha blending

The same way the obtained 3 rows of tiles are combined into a vessel mask of the whole image. After that the obtained mask is stretched to the size of the original image.

The results of segmentation model for vascular network on training set and validation set are shown in Fig. 4.

IV. EfficientNetB6 Classifier (without attention mask)

We examined a baseline image classification pipeline trained to categorize retinal fundus images into one of seven diagnostic classes. The model was trained on a curated version of the ODIR-5K dataset. Each fundus image (left or right eye) was labeled into one of the following categories:

- Pathological Myopia
- Hypertensive Retinopathy
- Diabetic Retinopathy
- Glaucoma
- Cataract
- Age-related Macular Degeneration (AMD)
- Normal (Healthy)

Each class was sampled with up to 250 left-eye and 250 right-eye images. Images were loaded, resized to 224×224, and paired with integer labels.

a) Data Augmentation.: Training images were augmented using the `ImageDataGenerator` utility with the following transformations:

- Rotation: $\pm 30^\circ$
- Width and height shift: 10%
- Zoom: 20%
- Horizontal flipping

These augmentations were applied to improve generalization and reduce the risk of overfitting.

b) Model Architecture.: The core of the model is `EfficientNetB6` [10] with `include_top=False`. Pretrained weights were used to initialize the base. A custom classification head was added:

- `GlobalAveragePooling2D`
- `Dense(224, activation='relu')`
- `Dropout(0.3)`
- `Dense(7, activation='softmax')` for 7-class prediction

c) Training Configuration:

- Loss function: `sparse_categorical_crossentropy`
- Optimizer: Adam with a learning rate of 1×10^{-4}
- Epochs: 30
- Batch size: 8
- Early stopping: Enabled (patience = 5)

d) Baseline result: Figure X shows the confusion matrix of the model on the test set, highlighting per-class prediction accuracy and common misclassification patterns (Fig. 5). It reveals both the strengths and limitations of the baseline model when distinguishing between retinal diseases.

Key Metrics:

- Overall test accuracy: 78.3%
- Highest per-class accuracy:
  - Myopia: 94%
  - Age-related Macular Degeneration (AMD): 90%
- Lowest per-class accuracy:
  - Diabetic Retinopathy: 29%
  - Healthy: 69%

V. EFFICIENTNETB6 CLASSIFIER using semantic attention mask

To improve classification performance, we introduced an attention mechanism that utilizes semantic segmentation masks generated for each fundus image. The core idea is to guide the classifier's focus toward clinically relevant anatomical regions—namely the optic disc, optic cup, and retinal vessels—by assigning higher weights to these structures and attenuating the influence of the background.

A. Attention Mask Generation

For each image, we applied pretrained segmentation models to generate binary masks corresponding to:

- Optic disc (channel 0),
- Optic cup (channel 1),
- Retinal vessels (channel 2).

These masks were resized to match the input image dimensions and normalized to the range [0, 1]. Each region was assigned a scalar weight based on its diagnostic relevance:

- $w_{\text{disc}} = 0.9$,
- $w_{\text{cup}} = 1.0$,
- $w_{\text{vessels}} = 0.8$,
- $w_{\text{background}} = 0.2$.

The values were selected empirically through ablation studies and reflect the clinical importance of each anatomical structure.

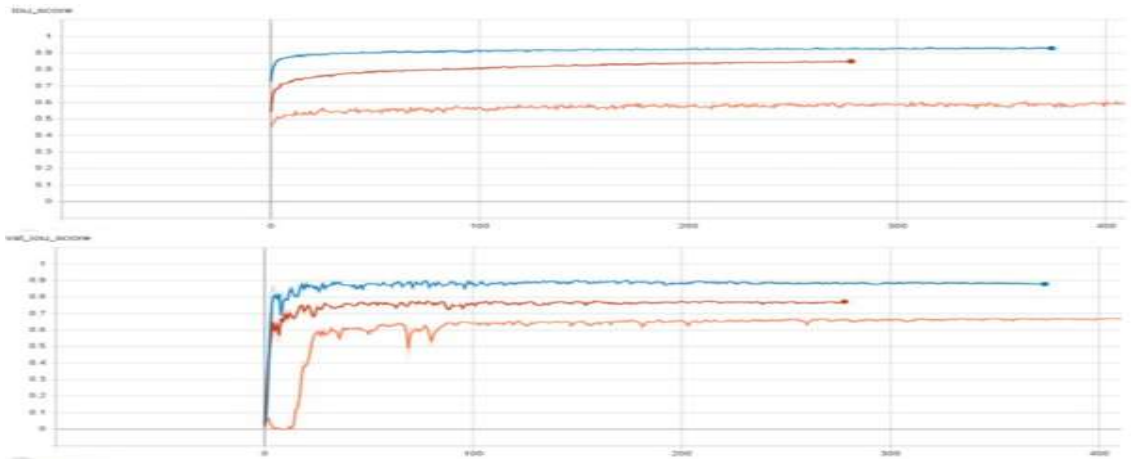

Figure 4. IoU score. first – training, second – validation. Blue – disk segmentation, red – excavation zone, orange – blood vessels

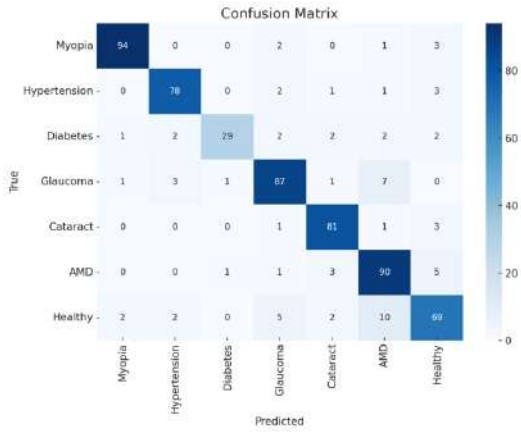

Figure 5. Confusion matrix for the baseline model.

B. Mask-Based Image Enhancement

A composite attention mask M is computed as a weighted sum of the individual segmentation maps:

$$M = w_{\text{disc}} \cdot \text{mask}_{\text{disc}} + w_{\text{cup}} \cdot \text{mask}_{\text{cup}} + w_{\text{vessels}} \cdot \text{mask}_{\text{vessels}} + w_{\text{background}} \cdot (1 - \text{combined\_mask})$$

where combined_mask denotes the union of all binary object masks. The attention-enhanced image I_{attn} is then obtained via element-wise multiplication:

$$I_{attn}(x, y) = I(x, y) \cdot M(x, y)$$

This operation suppresses less informative background regions while amplifying features in diagnostically critical areas.

C. Classification Pipeline

The modified images I_{attn} are fed into the same EfficientNetB6 classifier described earlier. The training configuration—loss function, optimizer, learning rate, and number of epochs—remains unchanged to ensure a consistent comparison with the baseline model.

This strategy allows the network to concentrate on regions most likely to contain pathological changes, and leads to measurable improvements in classification accuracy, especially in complex or borderline cases.

a) Improved result: After training model with attention masks the confusion matrix of the model shows optimized classification performance. The updated matrix demonstrates better per-class accuracy and reduced confusion among visually similar diseases compared to the baseline model (Fig. 6).

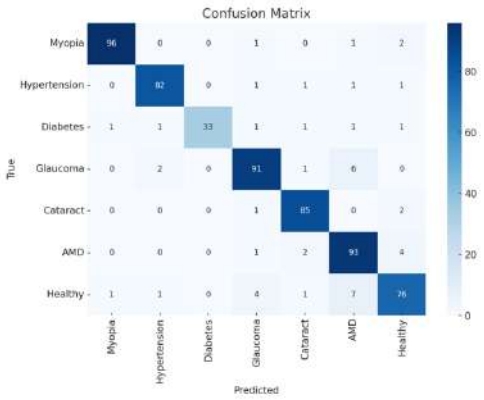

Figure 6. Confusion matrix for the improved model.

Key Metrics:

- Overall test accuracy: 92.2%
- Highest per-class accuracy:
  - Cataract: 96.6%
  - Myopia: 96.0%
- Lowest per-class accuracy:
  - Diabetic Retinopathy: 84.6%
  - Healthy: 84.4%

VI. Discussion

Within the framework of OSTIS (Open Semantic Technology for Intelligent Systems), attention masks and reflexivity become key components for building an intelligent diagnostic system based on the analysis of fundus images. At the core of the method is the use of semantic segmentation to highlight key anatomical structures — the optic disc, optic cup, and blood vessels — which are then used as attention masks in the EfficientNetB6 neural network, significantly improving classification accuracy from 78.3% to 92.2%. This approach is particularly valuable for diagnosing complex diseases such as diabetic retinopathy, where precise detail recognition is critically important. Within OSTIS, attention masks become a tool that implements the principle of reflexivity: the system can analyze its own decisions, adjusting the weights of these masks based on classification errors, which enhances its adaptability and effectiveness.

Attention masks are weight maps that highlight anatomically significant areas of the image, such as the optic disc, optic cup, and blood vessels. They are created using semantic segmentation and integrated into the OSTIS system as a tool that directs the classifier's focus to critically important zones. This allows the system to amplify signals from key areas, improving the accuracy of diagnosing pathologies such as glaucoma or diabetic retinopathy. For example, by applying attention masks, the model can more clearly distinguish subtle changes in eye structures, which is especially important for diseases with minor visual manifestations.

Representing knowledge about diseases and anatomy in a universal format ensures their reusability and scalability, allowing for the creation of flexible systems. This integration not only improves diagnostic accuracy but also makes the system capable of explaining its conclusions, which is crucial for medical practice where the interpretability of decisions plays a key role.

The integration of attention masks and reflexivity makes the OSTIS system not only more accurate but also interpretable. Attention masks help explain which areas of the image the classifier relies on, while reflexivity ensures transparency in the learning and adjustment process. This is particularly valuable in medical practice, where doctors need justification for diagnostic decisions.

VII. Conclusion

This study demonstrates that incorporating semantic segmentation into the classification pipeline significantly enhances the performance of retinal disease diagnosis from fundus images. By generating semantic masks of the optic disc, optic cup, and retinal vessels, and using them to guide the classifier's attention, we were able to improve both accuracy and robustness of the model.

The proposed attention mechanism allows the network to prioritize clinically relevant regions while suppressing less informative background areas. This research shows that semantic information, when embedded in the form of weighted attention masks, leads to better feature representation and higher classification performance, particularly in cases involving subtle or overlapping pathological signs.

Acknowledgment

The authors would like to thank the group of Laboratory of Information and Computer Technologies of Belarusian State Medical University for their valuable comments and help, in particular, Grigory Karapetyan and Kosik Ivan.

References

[1] Gurevich I. B, Yashina V. V., Ablameyko S. V., Nedzved A. M., Ospanov A. M., Tleubaev A. T., Fedorov A. A., Fedoruk N. A. Development and experimental investigation of mathematical methods for automating the diagnostics and analysis of ophthalmological images [Electronic resource]. *Pattern Recognit Image Anal*, 2018, № 28(4), pp. 612–636.

[2] Sarki, R. [et al.] Automatic detection of diabetic eye disease through deep learning using fundus images : a survey, IEEE Access, 2020, Vol. 8, P. 151133–151149, DOI: 10.1109/ACCESS.2020.3015258.

[3] Ting, D. [et al.] Artificial intelligence and deep learning in ophthalmology, Br J Ophthalmol, 2019, Vol. 103, P. 67–75, DOI: 10.1136/bjophthalmol-2018-313173.

[4] Ambika Selvakumar. Understanding Optic Disc Pallor – Shades of White. Mode of access: https://www.eophtha.com/posts/understanding-optic-disc-pallor-shades-of-white (accessed 24, Jan)

[5] Starovoitov V.V., Golub Y.I., Lukashevich M.M. Digital fundus image quality assessment. *System analysis and applied information science*, 2021, Volume 4, pp. 25- 38.

[6] Paperswithcode. Available at: https://paperswithcode.com/dataset/drive

[7] Paperswithcode. Available at: https://paperswithcode.com/dataset/chase-db1

[8] Paperswithcode. Available at: https://paperswithcode.com/dataset/hrf

[9] Adrian Rosebrock,Deep Learning for Computer Vision,PyImageSearch, 2017, 330 p.

[10] EfficientNet paper – Tan, M., Le, Q. v. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. 36th International Conference on Machine Learning, ICML 2019, 2019-June.

УЛУЧШЕНИЕ ДИАГНОСТИКИ ИЗОБРАЖЕНИЙ ГЛАЗНОГО ДНА С ИСПОЛЬЗОВАНИЕМ МАСКИ АЛГОРИТМА ВНИМАНИЯ НА ОСНОВЕ СЕМАНТИЧЕСКОЙ СЕГМЕНТАЦИИ

Гимбицкая Е. В., Свистунова К. И., Карапетян Г. М., Недзьведь А. М., Абламейко С. В.

В данной работе предлагается метод классификации офтальмологических заболеваний по изображениям глазного дна с применением семантической сегментации в качестве механизма внимания. В отличие от традиционных подходов, использующих всю область сетчатки, предложенная система акцентирует внимание на анатомически значимых зонах, выделенных посредством сегментации диска зрительного нерва, оптической чаши и сосудистой сети. Полученные сегментационные маски интегрируются в алгоритм классификации для улучшения извлечения признаков. В качестве классификатора используется модель на базе EfficientNetB6, позволяющая оценить эффективность предложенной стратегии. Результаты экспериментов демонстрируют повышение точности классификации по ряду метрик и успешность данного метода.

Received 23.03.2025