Segmenting the Optic Disc Using a Deep Learning Ensemble Model Based on OWA Operators

Mohammed Yousef Salem ALI [a,1], Mohamed ABDEL-NASSER [a,c]
Mohammed JABREEL [b] Aida VALLS [a] and Marc BAGET [d]

[a] *Departament Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Tarragona (Catalonia), Spain*
[b] *Microsoft Advanced Technology Lab, Cairo, Egypt*
[c] *Department of Electrical Engineering, Aswan University, 81528 Aswan, Egypt*
[d] *IISPV, Hospital Universitari Sant Joan de Reus, Spain*

Abstract. The optic disc (OD) is the point where the retinal vessels begin. OD carries essential information linked to Diabetic Retinopathy and glaucoma that may cause vision loss. Therefore, accurate segmentation of the optic disc from eye fundus images is essential to develop efficient automated DR and glaucoma detection systems. This paper presents a deep learning-based system for OD segmentation based on an ensemble of efficient semantic segmentation models for medical image segmentation. The aggregation of the different DL models was performed with the ordered weighted averaging (OWA) operators. We proposed the use of a dynamically generated set of weights that can give a different contribution to the models according to their performance during the segmentation of OD in the eye fundus images. The effectiveness of the proposed system was assessed on a fundus image dataset collected from the Hospital Sant Joan de Reus. We obtained Jaccard, Dice, Precision, and Recall scores of 95.40, 95.10, 96.70, and 93.90%, respectively.

Keywords. Fundus image, Optic disc segmentation, Deep learning, OWA operators

1. Introduction

Retinal diseases like diabetic retinopathy (DR) and glaucoma highly affect the optic disc (OD) of the human eye–OD is also called the optic nerve head. The early detection and treatment of DR and glaucoma are essential to prevent total vision loss [1]. Ophthalmologists inspect eye fundus images to detect the signs of such diseases. Yet, the manual analysis of many fundus images is expensive in terms of both time and effort. Modern computer-aided diagnosis (CAD) systems can analyze the medical images and provide a diagnosis as accurate as ophthalmologists with many years of experience [2]. Deep learning (DL) technologies have become the cornerstone of several modern CAD systems.

---

[1]Corresponding Author: Mohammed Yousef Salem Ali. E-mail: horbio10@gmail.com

Recently, several DL-based automated systems have been proposed to segment the OD automatically. Most of them use convolutional neural networks (CNNs) to automatically learn representative and high-level features from the input fundus images to achieve accurate segmentation [3]. For instance, the authors of [4] modified the U-Net model by reducing the size of the filters in all CNN layers to segment the OD directly. They stated that this modification produced a lighter model and improved the segmentation performance. With the DRIONS-DB dataset, they achieved IOU and Dice scores of 89 and 94%, respectively. With RIM-ONE v.3, they obtained IOU and Dice scores 89 and 95%, respectively. The authors of [5] modified the U-Net and DeepLabv3+ models by inserting an attention module between the encoder and decoder networks and adding a conditional random field layer in the output layer. The method was evaluated using DRIONS-DB, RIM-ONE v.3, and DRISHTI-GS fundus image datasets. They achieved a 95% Dice and 91% Jaccard with DRIONS-DB, a 97% Dice and a 94% Jaccard with RIM-ONE, and a 96% Dice and a 92%Jaccard with DRISHTI-GS.

In [6], a multi-label deep neural network called M-Net was proposed for segmenting OD and optic cup (OC) jointly. M-Net includes a multi-scale input layer, U-shape CNN, a side-output layer, and a multi-label loss function based on the dice loss. The side-output layer works as an initial classifier to provide a local prediction map for different scale layers. M-Net achieved an accuracy of 98.30% with the ORIGA dataset. The study confirmed that the accurate segmentation process of OD and OC jointly is essential to glaucoma detection and diagnosis. The authors of [7] leveraged the multi-task learning and dense connections to segment the OD and OC jointly. They evaluated their work on three datasets: DRISHTI-GS, RIM-ONE, and REFUGE. With the OD segmentation task, they achieved a 91.83% Jaccard and a 95.97% Dice with DRISHTI-GS, a 91.01% Jaccard and a 95.82% Dice with RIM-ONE, and an 88.37% Jaccard with REFUGE.

Although the methods mentioned above achieved good OD segmentation accuracy, no individual model performs best under all conditions. The fusion of the predictions of individual models can improve the segmentation accuracy [8]. In [9], a CNN model was trained using probability masks instead of binary masks to segment OD and OC. Each probability mask was obtained by fusing segmentation masks made by multiple experts.

In [10], an ensemble of different CNN architectures for medical image classification is used. The study demonstrated that different CNNs learn different levels of semantic image representation, and therefore an ensemble of CNNs produces richer features. [11] fused the predicted labels of five CNNs models (Resnet50, Inceptionv3, Xception, Dense121, and Dense169) to improve the results of DR classification in fundus images.

However, most existing methods use fixed weights associated with each model to construct the ensemble. This strategy may not be appropriate for all situations. Differently, in this paper we employ the ordered weighted averaging operator (OWA) [12] to aggregate the information provided by several OD segmentation models. More specifically, this paper proposes an ensemble-based OD segmentation system based on a set of deep learning models and an aggregation stage based on OWA.

OWA is an aggregation operator that uses a set of weights to define the aggregation policy for the inputs, which can vary from full conjunctiveness to full disjunctiveness. These weights do not indicate the importance of each source of information but the importance of the input values that are merged. The definition of the proper weights is a key issue in this aggregation operator. A dynamic way of generating weights is through

quantifier functions [12]. Notably, this OWA-based ensemble of DL models will allow the construction of a dynamic OD segmentation system.

The remainder of this paper is organized as follows. Section 2 explains the proposed OD segmentation system. Section 3 presents the experiments and results. Finally, the conclusion and future work are provided in section 4.

2. The Proposed Optic Disc Segmentation Method

To develop a reliable and efficient OD segmentation system, we first train several deep learning segmentation models. Figure 1 shows the procedure of segmenting a new image. The test image fed into *N* individual OD segmentation models. Then, the OWA operator aggregates the different segmentation masks given by the models to produce the final segmentation mask (final prediction).
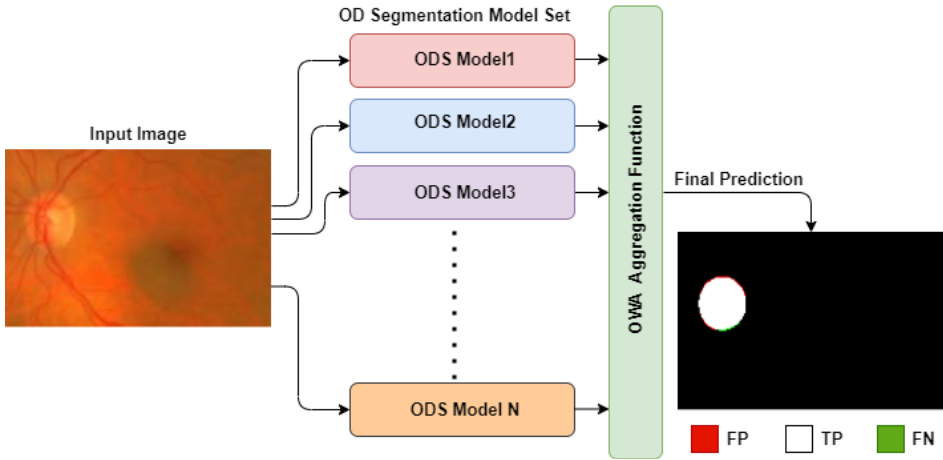

Figure 1. Structure of the proposed method.

2.1. Constructing individual OD segmantation models

In this study, we developed ten OD segmentation models based on different state-of-the-art deep learning-based semantic segmentation models. Specifically, we used the Unet, Gated Skip Connections (GSCs), DoubleU-Net, DeepLabV3+, CGNet, ERFNet, SegNet, ESNet, LinkNet, and SQNet models. Each of them was trained to maximize the segmentation accuracy. The architecture of all models based on the encoding and decoding method for work except the CGNet. The binary cross-entropy loss function and ADAM optimizer were used to train all models. The source code of CGNet, ERFNet, SegNet, ESNet, LinkNet, and SQNet can be found at `https://github.com/xiaoyufenfei/Efficient-Segmentation-Networks/blob/master/model.` We briefly introduce each model below.
- **UNet [13]** had five blocks in the contracting path (downsampling path) and five blocks in expanding path (upsampling path). The model contained a total of

twenty-three convolutional layers. Source code available at `https://github.com/jakeret/unet/blob/master/src/unet/unet.py`

- **Gated Skip Connections (GSCs) [14]** is a modified version of UNet. It had five encoder blocks with two convolutional layers, five decoder blocks with four convolutional layers, a total of thirty-three convolutional layers, and a gated skip connection mechanism in each decoder block with a convolutional layer.
- **DoubleU-Net [15]** included two stacked U-Net [13] in sequence. The first stack based on pre-trained VGG19 of an encoder with four blocks in each encoder-decoder, Second stack similar Unet with four blocks encoder-decoder, atrous spatial pyramid pooling after per encoder for each stack. Each block contained two convolutional layers and max-pooling. Source code available at `https://github.com/DebeshJha/2020-CBMS-DoubleU-Net`
- **DeepLabV3+ [16]** included ResNet50 backbone, deep separation convolution, and atrous separable convolution. Source code available at `https://github.com/srihari-humbarwadi/DeepLabV3_Plus-Tensorflow2.0/blob/master/deeplab.py`
- **CGNet [17]** included three down-sampling stages with Context Guided (CG) blocks for each stage. In total, it contained a total of fifty-one convolutional layers with no decoder modules.
- **ERFNet [18]** included two types of residual layer design (bottleneck and non-bottleneck), skip connections between encoder and decoder, contained a total of twenty-three convolutional layers.
- **SegNet [19]** had four blocks in the encoder part with thirteen convolutional layers of the VGG16 backbone network, four blocks in the decoder part with thirteen convolutional layers too.
- **ESNet [20]** had three blocks in the each encoder and decoder part,contained factorized convolution unit and parallel factorized convolution unit, contained a total of eighteen convolutional layers.
- **LinkNet [21]** had four blocks in each encoder and decoder part, four convolutional layers per each encoder block, residual block per two layers in each encoder block and two convolutional layers per each decoder block (26 layers).
- **SQNet [22]** is based on a modified SqueezeNet 1.1 (fire module). It contained eight SqueezeNet with three convolutional operations in each fire module, four parallel dilated convolutions and a refinement module in the decoder network. Exponential Linear Units were used in SQNet to avoid the bias shift.

2.2. OWA-based Aggregation Function

For the aggregation step, we propose the use of the Ordered Weighted Average operator defined by Yager [12]. This operator performs a weighted average of a vector of input values $X = (x_1, x_2, ... x_m)$, but before averaging, the values are sorted from the best to the worst (denoted with σ). In that way, the weighting vector W is associating importance to each position in the ordered input vector, regardless of the source of the value in that position. This is expressed in (Equation 1). Notice that the sum of weights must be equal to 1.

$$F_W(x_1, x_2, \ldots, x_m) = \sum_{k=1}^{m} w_k x_{\sigma_k} \tag{1}$$

The key property of this operator is that the weights can model different aggregation policies, ranging from situations of full andness (when $w_m = 1$ and rest are 0) to full orness (when $w_1 = 1$ and rest are 0). Any possibility in between can be represented with an appropriate combination of the weights.

Manually setting the weights for a certain policy may be difficult. One usual way of generating the weighting vector automatically is through Regular Increasing Monotone fuzzy linguistic quantifiers [12]. Each weight is then obtained using the expression in (Equation 2), where $Q(z)$ is a function that corresponds to a fuzzy quantifier. This quantifier can generate weights representing different "quantities" of agreement, which is a factor to determine the degree of andness/orness of the aggregation.

$$w_k = Q\left(\frac{k}{m}\right) - Q\left(\frac{k-1}{m}\right) \tag{2}$$

Three common aggregation strategies that can be defined using a fuzzy quantifier are the following: i) *At least half*, Q^1, which is more tolerant and focuses only on a small subset of the best values aggregated (Equation 3). It corresponds to a situation of low andness. ii) *As many as possible*, Q^2, which is quite strict in the aggregation as it gives most importance to the worst values aggregated (Equation 4). It represents a situation of high andness. iii) *Most*, Q^3, which represents a case of finding the majority value (Equation 5). It corresponds to a situation of moderate andness (close to neutrality).

$$Q^1(z) = \begin{cases} 2z & \text{if } 0 \le z \le 0.5 \\ 1 & \text{if } 0.5 < z \le 1 \end{cases} \tag{3}$$

$$Q^2(z) = \begin{cases} 0 & \text{if } 0 \le z \le 0.5 \\ 2z - 1.0 & \text{if } 0.5 < z \le 1 \end{cases} \tag{4}$$

$$Q^3(z) = \begin{cases} 0 & \text{if } 0 \le z \le 0.3 \\ 2(z - 0.3) & \text{if } 0.3 < z \le 0.8 \\ 1 & \text{if } 0.8 < z \le 1 \end{cases} \tag{5}$$

These three quantifiers are the ones that have been used for aggregating the probability values obtained with the different DL models for each class. The aggregation is applied at each pixel of the image studied. A threshold of 0.5 is applied to the result and the class with the maximum activation is taken as label.

2.3. Evaluation Metrics

The evaluation is done with the segmented masks and the corresponding ground truth. The segmented masks were binarized so that pixels in the OD have the label '1' while pixels in the background have the label '0'. We computed the number of true positive (TP), true negative (TN), false positive (FP), and false-negative (FN) pixels.

Then, we computed the evaluation metrics commonly used in the OD segmentation task: the intersection-over-union (IOU), Precision (Prec.), Recall, and Dice coefficient (Dice).

i) *IOU* is the ratio of the intersection between the two masks (original and predicted mask) for OD with respect to their union. It is computed using Equation 6.

$$IOU(A,B) = \frac{A \cap B}{A \cup B} = \frac{TP}{TP + FP + FN} \tag{6}$$

ii) *Precision* is the ratio of TP and the pixels predicted as OD.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

iii) *Recall* is the ratio of TP and the correct pixels in OD according to ground truth.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

iv) *Dice coefficient* makes an harmonic mean of precision and recall. It is also referred to as F-score.

$$\text{Dice} = 2 \cdot \frac{\text{Precision. Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

3. Results and Discussions

3.1. Dataset and Experimental Setup

Here, we used an in-house fundus image dataset collected from Hospital Universitari Sant Joan de Reus (SJR). It includes 105 images that were manually labeled by ophthalmologists in SJR. The original size of the images is 3008×2000. We randomly divided the dataset into three subsets: the training set (56 images), the validation set (14 images), and the testing set (35 images). We used data augmentation techniques to increase the training data set size, generating a training set with 1120 images. Then, we resized the images and the ground truths to 256×384 to make the training process faster. The validation set was used to save the best checkpoint of the trained models. We employed the binary cross-entropy loss function in all models. We trained the models with 50 epochs using Adam optimizer with a learning rate of 0.001, and a batch size of 4 images.

3.2. Analysis of the Results of Individual DL models

We trained and tested 10 state-of-the-art deep CNN independently for OD image segmentation on the SJR dataset. The results are shown in Table 1. We noted that the GSCs model has the best result of IOU and Dice. Then ERFNet gives the best result of the Precision metric. Also, the DoubleU-Net has the best results of the Recall metric. So, we can see that there is not a unique winner method in all quality indices. This fact motivates the idea of using an ensemble of a subset of these models. To illustrate the results, we displayed some sample masks obtained for the three best IOUs OD segmentation models

Table 1. Performance comparison on SJR dataset.

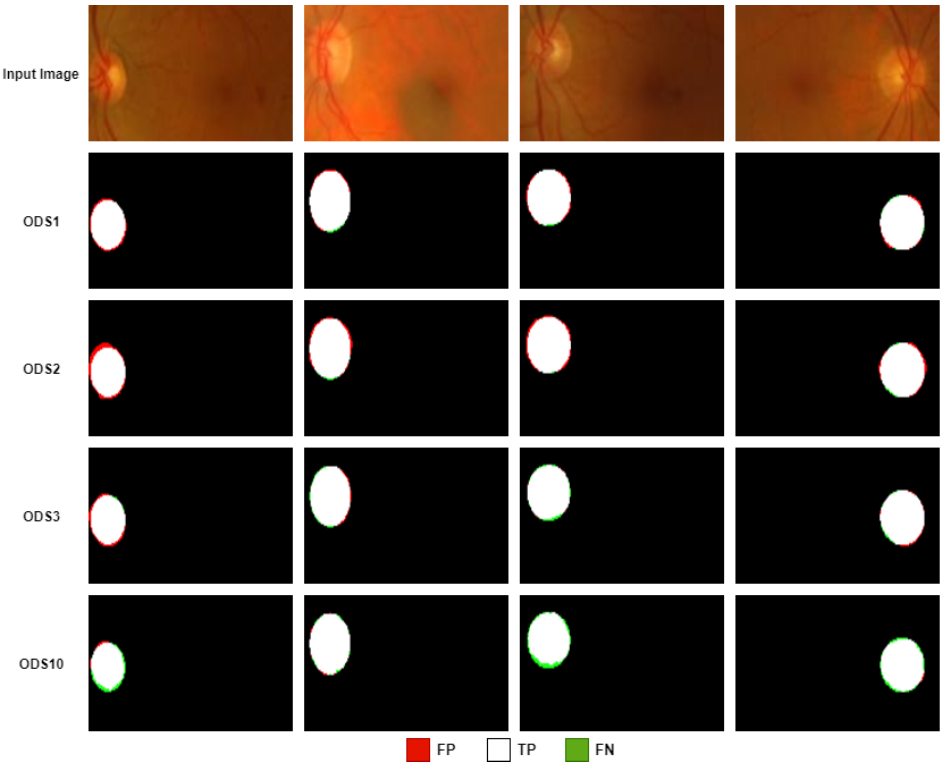| Models	Evaluation Metrics			
	IOU (%)	Dice (%)	Prec.(%)	Recall (%)
ODS1: GSCs	**95.1**	**94.9**	97.0	92.9
ODS2: DoubleU-Net	94.4	94.2	93.3	**95.0**
ODS3: DeepLabV3+	94.1	93.8	97.7	90.1
ODS4: U-Net	92.4	91.8	96.6	87.5
ODS5: LinkNet	92.4	91.8	99.7	85.3
ODS6: SQNet	91.9	91.2	97.6	84.8
ODS7: ESNet	91.9	91.2	99.5	84.3
ODS8: CGNet	91.6	90.8	99.6	83.7
ODS9: ERFNet	91.5	90.7	**99.8**	83.4
ODS10: SegNet	90.5	89.4	99.3	81.8

Figure 2. Results of the best three OD segmentation models and the worst one. Here ODS1, ODS2, ODS3, ODS10 indicate to GSCs, DoubleU-Net, DeepLabV3+, SegNet models.

and the worst one in Figure 2. Pixels in red are the FP, in green the FN, while white ones TP are the correctly classified.

3.3. Analysis of the Results of the OWA-based DL Ensemble System

In this section, we compared the three OWA aggregation policies defined before in

Section [2:](#) Q^1 uses a disjunctive approach, considering only the models with the best performance, Q^2 applies a strong conjunctive approach, requiring that most models agree on the same output, while Q^3 takes an intermediate point between the previous ones.

We used different subsets of models to build the ensemble. Reducing the number of models is interesting to decrease the execution time. After ranking the DL models with the IOU index, we compared a set with the Top-7 models: GSCs, DoubleU-Net, DeepLabV3+, U-Net, LinkNet, SQNet, and ESNet; a set with the Top-5 models: GSCs, DoubleU-Net, DeepLabV3+, U-Net, and LinkNet; and with the Top-3 models: GSCs, DoubleU-Net, and DeepLabV3+. Results are shown in Table [2.](#)

Table 2. Results of different OWA's policies with a different number of aggregated models.

Aggreg. Mod.	Evaluation Metrics of OWA Policies (%)											
	At least half, Q^1				As many as possible, Q^2				Most, Q^3			
	IOU	Dice	Pre	Rec	IOU	Dice	Pre	Rec	IOU	Dice	Pre	Rec
10 Mod.	93.4	92.9	**99.6**	87.3	93.7	93.2	**99.5**	87.9	93.5	93.0	**99.6**	87.5
Top-7	94.1	93.8	99.2	89.2	94.4	94.1	99.0	89.8	94.3	94.0	99.1	89.6
Top-5	94.9	94.6	98.3	91.4	95.1	94.8	98.3	91.9	95.0	94.7	98.2	91.7
Top-3	**95.2**	**95.0**	97.4	**93.0**	**95.4**	**95.1**	96.7	**93.9**	**95.3**	**95.1**	97.2	**93.3**

With the three aggregation policies, we can see that performance indicators are better when we reduce the number of models in the ensemble. The best results are obtained with the Top-3 ensemble, which achieved an IOU of 95.4% and Dice of 95.1%. We observe that the IOU and Dice scores of the aggregated models are better than the ones of the individual models. However, precision is higher when using more models. An improvement in recall means that we reduce the False Negatives. The non-improvement of precision means that the number of FPs is the same or increase a bit. However, as IOU and Dice are better it means that the improvement in FNs is considerably larger than the increase of FPs.

Regarding OWA, we can also observe that the most conjunctive policy (i.e. as many as possible, Q^2) is outperforming the others. Comparing the results with the Q^2 ensemble system and the individual models, we observe an improvement in the performance (Table [1](#)). In this experiment, results improve when using a strict conjunctive aggregation policy, which is the one that requires a high agreement of the models in order to choose the final class. As OWA is using the best performing subset of n models, this aggregation is taking into consideration the consensus of the n models regarding the class probability of each pixel (i.e. probability of each class given by each model). As this consensus may be difficult to find, it is reasonable that smaller subsets perform better than adding more models with fewer performance indicators. This observation is interesting because it means that it is not necessary to include a large number of DL models in the ensemble. From the improvement with respect to individual models, we can conclude that the ensemble approach can improve the segmentation result, correcting some of the mistakes of individual models. Figure [3](#) displays the segmentation made with the three different OWA policies when using the Top-3 methods with the same images than previous figure.

It is interesting to determine the statistical significance of the differences in performance between the proposed ensemble and GSCs (the best individual model) in terms of the IOU and Dice. To do so, we used Student's t-test (significance level < 0.05) to determine the difference in IOU and Dice values. The p-values obtained are higher than 0.05, indicating no statistical significance.
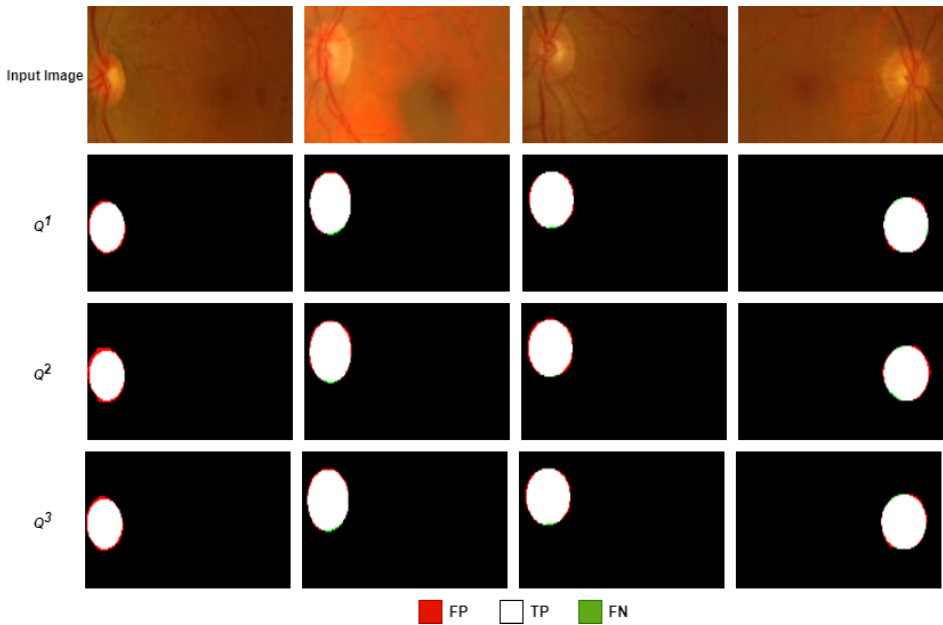
Figure 3. Results of the three policies of the OWA proposed method. Here Q1, Q2, Q3 indicate to the policies

4. Conclusions and future work

This paper presented an automated system for segmenting the OD in fundus images of the human eye based on accurate deep learning-based segmentation models and OWA operators. We assessed the performance of three different policies and a different number of individual OD segmentation models. We found that a high conjunctive aggregation with a reduced subset of the classifiers leads to the best OD segmentation model. With an in-house fundus images dataset collected from the Hospital Sant Joan de Reus, our system achieved an IoU and Dice scores higher than 95%, a precision higher than 96%, and a recall higher than 93%. Using an ensemble of DL models increases the training time significantly because several individual OD segmentation models must be built before constructing the ensemble. The fact that a combination of only three models has given the best results makes this ensemble approach feasible. The inference time is low enough to be used in a hospital in real-time. Moreover, the improvement achieved with the proposed ensemble system can give a more accurate segmentation of the OD, which reduces human errors for ophthalmologists in detecting eye pathologies. Future work will include using the proposed OD segmentation model to develop detection tools for eye diseases like glaucoma or diabetic retinopathy.

Acknowledgements

References

[1]   Mary VS, Rajsingh EB, Naik GR. Retinal fundus image analysis for diagnosis of glaucoma: a comprehensive survey. IEEE Access. 2016;.

[2]   Jani K, Srivastava R, Srivastava S, Anand A. Computer aided medical image analysis for capsule endoscopy using conventional machine learning and deep learning. In: 2019 7th International Conference on Smart Computing & Communications (ICSCC). IEEE; 2019. p. 1–5.

[3]   Fourcade A, Khonsari R. Deep learning in medical image analysis: A third eye for doctors. Journal of stomatology, oral and maxillofacial surgery. 2019;120(4):279–288.

[4]   Sevastopolsky A. Optic disc and cup segmentation methods for glaucoma detection with modification of U-Net convolutional neural network. Pattern Recognition and Image Analysis. 2017;27(3):618–624.

[5]   Bhatkalkar BJ, Reddy DR, Prabhu S, Bhandary SV. Improving the performance of convolutional neural network for the segmentation of optic disc in fundus images using attention gates and conditional random fields. IEEE Access. 2020;8:29299–29310.

[6]   Fu H, Cheng J, Xu Y, Wong DWK, Liu J, Cao X. Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. IEEE transactions on medical imaging. 2018;37(7):1597–1605.

[7]   Tabassum M, Khan TM, Arsalan M, Naqvi SS, Ahmed M, Madni HA, et al. CDED-Net: Joint segmentation of optic disc and optic cup for glaucoma screening. IEEE Access. 2020;8:102733–102747.

[8]   Dietterich TG. Ensemble Learning, The Handbook of Brain Theory and Neural Networks, MA Arbib. Cambridge, MA: MIT Press; 2002.

[9]   Mangipudi PS, Pandey HM, Choudhary A. Improved optic disc and cup segmentation in Glaucomatic images using deep learning architecture. Multimedia Tools and Applications. 2021;p. 1–21.

[10] Kumar A, Kim J, Lyndon D, Fulham M, Feng D. An ensemble of fine-tuned convolutional neural networks for medical image classification. IEEE journal of biomedical and health informatics. 2016;21(1):31–40.

[11] Qummar S, Khan FG, Shah S, Khan A, Shamshirband S, Rehman ZU, et al. A deep learning ensemble approach for diabetic retinopathy detection. IEEE Access. 2019;7:150530–150539.

[12] Yager RR. Quantifier guided aggregation using OWA operators. International Journal of Intelligent Systems. 1996;11(1):49–73.

[13] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer; 2015. p. 234–241.

[14] Jabreel M, Abdel-Nasser M. Promising crack segmentation method based on gated skip connection. Electronics Letters. 2020;56(10):493–495.

[15] Jha D, Riegler MA, Johansen D, Halvorsen P, Johansen HD. Doubleu-net: A deep convolutional neural network for medical image segmentation. In: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS). IEEE; 2020. p. 558–564.

[16] Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV); 2018. p. 801–818.

[17] Wu T, Tang S, Zhang R, Cao J, Zhang Y. Cgnet: A light-weight context guided network for semantic segmentation. IEEE Transactions on Image Processing. 2020;30:1169–1179.

[18] Romera E, Alvarez JM, Bergasa LM, Arroyo R. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. IEEE Transactions on Intelligent Transportation Systems. 2017;19(1):263–272.

[19] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence. 2017;39(12):2481–2495.

[20] Wang Y, Zhou Q, Xiong J, Wu X, Jin X. Esnet: An efficient symmetric network for real-time semantic segmentation. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). Springer; 2019. p. 41–52.

[21] Chaurasia A, Culurciello E. Linknet: Exploiting encoder representations for efficient semantic segmentation. In: 2017 IEEE Visual Communications and Image Processing (VCIP). IEEE; 2017. p. 1–4.

[22] Treml M, Arjona-Medina J, Unterthiner T, Durgesh R, Friedmann F, Schuberth P, et al. Speeding up semantic segmentation for autonomous driving. In: MLITS, NIPS Workshop. vol. 2; 2016. .