A Method based on Deep Neural Network for Instance Segmentation of Retinal Lesions caused by Diabetic Retinopathy

Carlos Santos[*][†], *Member, IEEE*, Marilton Aguiar[†], Daniel Welfer[‡],
Marcelo Silva[§], Alejandro Pereira[§], Marcelo Ribeiro[§], Bruno Belloni[¶]
[*]Computer Center, Federal Institute of Education, Science and Technology Farroupilha, Alegrete, Brazil
[†]Postgraduate Program in Computing, Federal University of Pelotas, Pelotas, Brazil
[‡]Postgraduate Program in Computer Science, Federal University of Santa Maria, Santa Maria, Brazil
[§]Federal University of Pelotas, Pelotas, Brazil
[¶]Federal Institute of Education, Science and Technology Sul-Rio-Grandense, Passo Fundo, Brazil
Email: carlos.santos@iffarroupilha.edu.br

***Abstract*—Diabetic Retinopathy is one of the main causes of vision loss and can be identified through ophthalmological exams that aim to locate the presence of retinal lesions such as microaneurysms, hemorrhages, soft exudates, and hard exudates. The development of computerized methods to perform the instance segmentation of lesions may support in the early diagnosis of the disease. However, the instance segmentation of retinal artifacts is a complex task due to factors such as the size of objects and their morphological characteristics. This article proposes a method based on a Mask R-CNN neural network architecture to perform instance segmentation of lesions associated with diabetic retinopathy. The proposed method was trained, adjusted, and tested using the public DDR and IDRiD Diabetic Retinopathy datasets, and implemented with the Detectron2 and OpenCV libraries. The proposed method reached in the DDR dataset, using the SGD optimizer, the mAP of 0.2660 for the limit of IoU of 0.5 in the validation step. The results obtained in the experiments demonstrate that the proposed method showed promising results in the instance segmentation of fundus lesions.**

***Index Terms*—fundus image, lesions detection, instance segmentation, Mask R-CNN**

I. Introduction

The development of solutions for automated medical image diagnosis is an expanding scientific research field. Digital medical images are present in most diagnostic laboratories, providing easy manipulation through various computerized systems. The analysis of biomedical images through resources extracted from public datasets, and the construction of methods based on deep learning algorithms provide more subsidies for the decision-making of specialist physicians during diagnosis [1]. The implementation of measures that guarantee the rapid diagnosis of diseases, as well as the implementation of preventive measures and effective treatment are fundamental [2]. Diabetic retinopathy (DR) is a disease that affects the eyes and is caused by diabetes, being one of the main causes of vision loss in adults of working age [3]. Vision loss resulting from DR can be prevented when treated early [4]. DR is usually identified through eye exams that aim to identify retinal lesions, including microaneurysms (MA), hemorrhages (HE), soft exudates (SE), and hard exudates (EX). However, early screening of HR using traditional methods is a challenge due to the scarcity of professionals and resources to meet the growing demands, especially in poorer regions [5]. A computerized method can assist in the process of identifying retinal lesions and aid in the diagnosis and screening of DR. In addition, only an automatic method is able to accurately quantify the increase or decrease of the retinopathy lesions when applied to temporal images of the same patient. In the literature, solutions based on two-stage detectors have been presented to assist in the identification of the disease, but there are still limitations in the results presented by these works, mainly in the precision associated with the identification of very small objects in the fundus images.

In this context, the motivation for this article is to present a new method for instance segmentation of retinal lesions associated with DR, and thus assist in the identification and diagnosis of the disease. Instance segmentation is a hybrid of object detection and image segmentation, where *pixels* are not only classified according to the class they belong to, but individual objects within those classes are also extracted [1], [6], which is important when it comes to medical imaging. The detection with instance segmentation of fundus lesions is still a little explored problem and with limited results. In this context, the main contribution of this work is to present a new method based on deep learning to perform instance segmentation of retinal lesions.

The article is structured as follows: Section II describes related works. In Section III the materials and methods used for this work are presented. Section IV will describe the results and discussions obtained through the proposed method. Finally, in Section V the final considerations will be described.

II. Related Work

According to the challenges discussed in the previous Section, the following methods proposed in the literature that aim

979-8-3503-2028-2/22/$31.00 ©2022 IEEE
DOI 10.1109/CSCI58124.2022.00062

to instance segmentation of fundus lesions will be reviewed. Li et al. [7] presented a new diabetic retinopathy dataset called Dataset for Diabetic Retinopathy (DDR), and evaluated deep learning models for the classification, detection, and segmentation of retinal lesions. The results presented by authors in the semantic segmentation of microlesions, as in the case of microaneurysms, demonstrated the difficulty of the models used in identifying small objects in the fundus images because these lesions have only a few *pixels*.

The work by Dai et al. [8] presented a system to classify DR and detect fundus lesions. The study had some limitations: 1) was the exclusive use of a private DR dataset to perform the training of the deep learning models, using the public DR dataset Kaggle eyePACS only for the validation of the subnet responsible for the classification of DR; and, 2) the subnet that detects the lesions was tested only in the private dataset used by the authors due to the absence of fundus lesion annotations in the public dataset used in the experiments.

Shenavarmasouleh et al. [9] propose an architecture for detecting fundus lesions. The proposed work was limited to performing only the detection of exudates and microaneurysms. As future work, the authors intend to create an architecture capable of performing the instance segmentation of fundus lesions.

Although deep learning has the potential to solve tasks associated with medical imaging, there are still open questions and limitations in the results presented. To mitigate this limitations, this work intends to present an method capable of performing the segmentation of fundus lesions with the support of image pre-processing techniques and data augmentation together with a deep neural network implemented based on a Mask R-CNN pre-trained to improve the accuracy of identification of retinal lesions associated with DR.

III. Materials and Methods

The method was developed based on the Mask R-CNN [10] architecture, as illustrated in the block diagram shown in Fig. 1. For the construction of the architecture, we used the open-source library Detectron2 [1], [11], [12]. To carry out the experiments, a microcomputer with a Core i7-10700F @16x 2.90 GHz processor, with 16 GB of RAM, and a 12 GB NVIDIA RTX 3060 VRAM GPU was used.

The Mask R-CNN architecture is a model capable of detecting and segmenting object instances. This model extends the Faster R-CNN [10] object detection architecture, adding a parallel framework to predict object segmentation masks. Instance segmentation combines object detection tasks, where the objective is to classify and locate objects individually using a bounding box and also to locate each *pixel* of each detected object in the image. This architecture works in two stages. The first one consists of using a Region Proposal Network (RPN) [13], [14] to select the bounding boxes (BBox) of candidate objects. The second one aims to classify the candidate boxes, refine the boxes and predict the masks of the objects (Mask). Models that perform object detection, such as Faster R-CNN [13], SSD [15] and YOLO [16], draw a bounding box around detected objects, while Mask R-CNN provides instances segmentation in *pixels* for each object located in the image.

In the case of object detection, there is the possibility of lesions having their bounding boxes detected overlapping, making it difficult to visualize these lesions, and consequently the diagnosis. With instance segmentation, it is possible to detect the lesions and also to know the locations of the *pixels* of each lesion (borders), and consequently, to verify the exact size and extension of each detected lesion. In addition, two-stage detectors are often more accurate in locating and classifying objects than single-stage detectors in detecting small objects [17], [18], especially when they appear clustered in the image [19]. In two-stage detectors, there is a stage for the identification of a subset of regions in an image that may contain an object, and a second stage is used to perform the classification of the object in each region.

With the proposed method for instance segmentation, it was possible to perform a more accurate detection of fundus lesions, as well as a more efficient way to provide the specialist with better visualization of the exact extent of each detected lesion and assist him in the diagnosis of the disease. Below, we detail each part of the adopted methodology.

A. Dataset and Pre-Processing of Images

Two public datasets with fundus images were used to carry out the experiments, among which the dataset DDR [7], which has 757 images, and the dataset IDRiD [20], which has 81 images. Both datasets have *Ground Truth* for lesions MA, HE, SE and EX at the *pixel* level. We use the MS COCO annotation format, in which object annotations in the form of bounding boxes and object polygons are stored in a file in JavaScript Object Notation. The datasets have different characteristics from each other, such as the quantity and quality of the images, the annotation method and quantity of lesion annotations and the availability of *Ground Truth*. To perform the process of creating the annotations of the lesions in the form of polygons, the image files together with the binary masks of the lesions provided by the DDR dataset were used to capture the contour of the lesions through the function `find_contours()` from OpenCV. After identifying the contour of the lesions, the annotations were created with the help of the `create_annotation_format()` function. Finally, these annotations are transformed for the standard COCO JSON using the `get_coco_json_format()` function for training of the proposed method. Fig. 2 presents an example of a fundus image of the DDR dataset in (a), and the same image with the lesion annotations in the form of bounding boxes and the generated annotations in polygon format in (b).

As the works by Santos et al. [21] and Alyoubi et al. [22], we performed a pre-processing step for partial *cropping* of the black background of retinal images. To operate partial removal of the black background, the Hough Transform (HT) [23] was used. First, pre-processing was performed through the application of the median filter technique to smooth the images

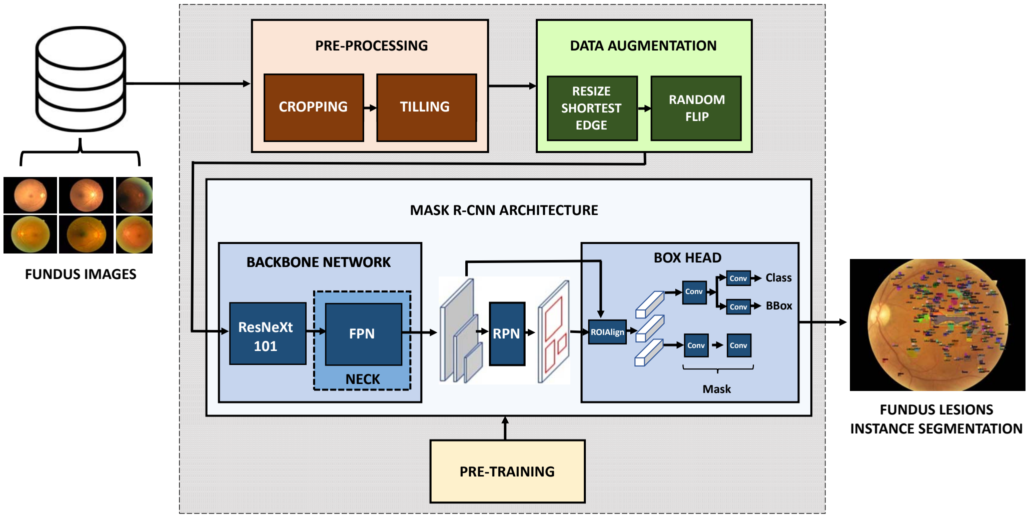

Fig. 1. Block diagram of the proposed method for instance segmentation of lesions fundus. First, the images are passed to the Pre-processing block for partial elimination of the black background of the images (Cropping) and the creation of sub-blocks of the images (Tilling). Then, the pre-processed images are transferred to the Data Augmentation block and next for training the proposed method. But, before, a pre-training step is performed with the weights fitted to the COCO dataset.

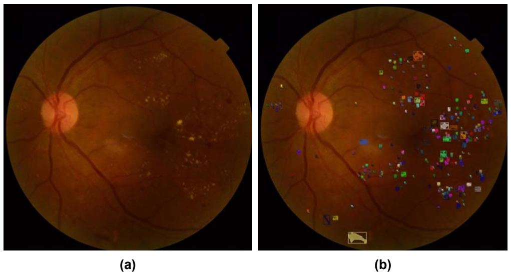

Fig. 2. DDR dataset fundus image (a); and, (b) annotations in the format of bounding boxes and polygons for training the deep neural network that composes the proposed method. As in instance segmentation the classification is performed at the *pixel* level, each instance of a lesion class was labeled with a different color.

and eliminate irrelevant details for the detection of the retinal circumference. Next, was necessary to threshold the images, followed by the detection of edges using the Sobel filter. After locating the retina, we transform its circumference into its equivalent rectangle to partially remove the black background.

The last pre-processing step is to perform *Tilling*. In fundus imaging, the detection of microlesions, as in the case of microaneurysms, remains a challenge. If the lesion area is not large enough, the signal propagated in the convolutional layers will be small while the model training is performed, leading to gradient dissipation. In addition, very small objects are more susceptible to data labeling errors, where accurate lesion identification can be impaired. As the work proposed by Santos et al. [21], the solution we adopted was the implementation of the *Tilling* method, in which the original images are cut into blocks (*tiles*). We create *tiles* of size 2×2. Each sub-image generated in this process remained with its respective lesions and annotations (*Ground Truth*), with no loss of information. After the application of *Tilling*, the resolution of the lesions present in the sub-images became higher than the non-partitioned images that had their resolutions reduced to be used in the input layer of the neural network. To minimize the risk of information loss from these lesions, we define an *overlap* area, in which each block will have an overlap area of 15% with its neighboring blocks. After the application of *Tilling*, we verified an increase in the precision of the proposed method due to better extraction of characteristics of fundus lesions.

B. Data Augmentation

The limited amount of labeled lesions available in public DR datasets restricts the amount of features extracted by the deep neural network during the training stage. In addition, microlesions often have gradient dissipation problems. Due to these problems, *on-the-fly* data augmentation was performed, in which the data loader applied the augmentation methods *Resize Shortest Edge* and *Random Flip* of the Detectron2 in the images of the public DR dataset used for training the model. *Resize Shortest Edge* resizes the image while keeping the aspect ratio unchanged, trying to scale the shortest edge to the given `short_edge_length` $(640, 672, 704, 736, 768, 800)$, as long as that the longest border does not exceed the defined `max_size` (1333). After performing the data augmentation, we trained the proposed method to perform the detection and segmentation of fundus lesions. Details about the architecture of the deep neural network are presented below.

C. Deep Neural Network Architecture

The proposed method performs the detection and segmentation at the *pixel* level of lesions present in fundus images.

First, the proposed region of the images is verified and classified. Then, bounding boxes and segmentation masks for the identified lesions are generated. The process of creating the mask of each lesion is performed using an additional convolutional neural network over a feature map, in which a matrix is generated and filled with 1 in all places where the *pixel* belongs to the lesion, and 0 as output in other locations.

The architecture of the neural network that composes the proposed method is basically constituted by three main modules, being a *Backbone*, a *Region Proposal Network* (RPN), and a *Box Head* (*ROI Head*). *Backbone* is a conventional convolutional neural network, and its purpose is to extract the characteristics of lesions in the fundus images. In Detectron2 the resolution of the input image does not need to be the same size as the input of the pre-trained model. Therefore, *a priori* it is possible to use any image resolution for *Backbone* input. However, due to limitations associated with the hardware equipment used to perform the experiments, we chose to use the standard Detectron2 resolution, which resizes the input images according to the parameters `INPUT.MIN_SIZE_TRAIN` equal to 800 *pixels*, and `INPUT.MAX_SIZE_TRAIN` equals 1333 *pixels*. As the images in the DDR dataset are wider than the maximum size defined above, Detectron2 resizes the fundus images to the width of 1333 *pixels*, adjusting the image height size proportionally to the width and according to the minimum defined size of 800 *pixels*. It is important to note that the images of the DDR set have variable sizes (height and width).

The *Backbone* of the architecture is composed of a residual network with convolutional layers clustered called ResNeXt [24]. In our experiments, the ResNeXt-101-32 × 8d-FPN architecture obtained the best precision in lesion detection and segmentation. ResNeXt consists of a structure that contains multiple *bottleneck* blocks. The ResNeXt-101-32 × 8d-FPN architecture is composed of 101 layers and cardinality (grouping) of convolutions of 32 groups and group width of 8 dimensions (88 million parameters). We adopted ResNeXt-101-32 × 8d because this architecture implements cardinality, which allowed us to improve the accuracy of lesion classification without, however, increasing the computational complexity of the architecture with the addition of parameters. The *Backbone* structure is followed by a *Feature Pyramid Network* (FPN) [25]. The FPN is used with the *Neck* of the architecture, being an extension of the deep neural network used in the *Backbone*, whose objective is to extract features and better represent the lesions at different scales. The FPN has five scales with outputs named P2, P3, P4, P5 and P6, respectively, with channel size $C =$ to 256 for all scales, and *stride* size $S = (4, 8, 16, 32, 64)$, respectively. Therefore, if only an input image of size 1333×1333 *pixels* is used in the input of *Backbone*, the sizes of the output feature maps of layers P2, P3, P4, P5 and P6 will be $334 \times 334 \times 256$, $166 \times 166 \times 256$, $84 \times 84 \times 256$, $40 \times 40 \times 256$ and $20 \times 20 \times 256$, respectively. Thus, layers P2 and P3 are used for detecting small objects, while layers P5 and P6 are responsible for detecting larger objects. FPN extracts feature maps at various scales with different receptive fields.

Next, the architecture has an RPN module, whose task is to inspect the entire FPN of the *Backbone* from top to bottom, to propose regions that may contain lesions in the fundus images. In Detectron2 all computation performed by the RPN is performed on the GPU. The RPN module uses anchors which are a set of boxes with predefined locations, where the anchors are sized according to the input images. Individual anchors are assigned to classes and bounding boxes. The RPN generates two outputs for each anchor: the anchor class and the bounding box specification. It should be noted that the RPN detects regions based on multi-scale features. By default, approximately $1,000$ cash proposals are obtained with confidence scores.

The last module of the architecture is the *Box Head*, responsible for cutting and interpolating the feature maps of the region proposals generated in the RPN. Only the characteristics of FPN layers P2, P3, P4, and P5 are used in *Box Head*. In addition, *Box Head* obtains the location of fitted boxes and sorting results through fully connected layers. This module has an ROI Pooling. According to [10], the selected feature map regions are misaligned with the proposed regions of the original image. As image segmentation requires specificity at the *pixel* level, this problem can cause inaccuracies during segmentation. To solve this problem a function from the Detectron2 ROIAlignV2 is used [10] so that the feature map is sampled at different points and then a bilinear interpolation is applied to obtain the precise position of the *pixel*.

After ROI Pooling, the cut features are used in the *Head* architecture. In the case of Mask R-CNN, there are two types of heads: *Box Head* and *Mask Head*. The calculation of the loss function of the outputs during training is performed using two functions: 1) *localization loss* (`loss_box_reg`), obtained through the function *Smooth L1 loss*; and, 2) *classification loss* (`loss_cls`), obtained through the cross entropy loss function *Softmax* [11]. The results of these *losses* are added to the losses calculated in the RPN (`loss_rpn_cls` and `loss_rpn_cls`), and added to the total loss [11]. To make the inferences of fundus lesions a post-processing step is performed to filter the low-scoring bounding boxes. For this, the technique of non-max suppression (NMS) is applied to eliminate ROIs that are below a pre-defined score threshold.

D. Pre-training

We use transfer learning to pre-train the neural network architecture. The pre-trained weights were imported into the COCO [26] dataset to initialize the weights of the neural network architecture used in the proposed method. We modified the output of the proposed model to suit the instance segmentation of retinal lesions associated with DR, preserving the weights of the upper layers. The training involve four main steps namely: 1) The initial layers of the architecture are pre-trained with the weights of the COCO dataset; 2) The last layers are cut and replaced with new layers; 3) The new layers added are adjusted in the DR dataset; and 4) After fine-tuning the final layers of the architecture, the entire neural network is retrained, so that small adjustments are made to the weights of the entire architecture.

E. Model Training and Adjustment

To carry out the training of the proposed method, images from the public datasets DDR and IDRiD were used. To carry out the experiments, we used the method of dividing the data sets into Training, Validation, and Testing in a proportion of 50:20:30, respectively. The fine-tuning of the proposed method aimed to optimize the hyperparameters to achieve more accurate results in the segmentation of instances of fundus lesions. The best fit of hyperparameters is shown in Table I.

TABLE I
ADJUSTED HYPERPARAMETERS OF THE PROPOSED METHOD IN THE VALIDATION STEP USING THE DDR DATASET.

Hyperparameter	**Value**
Number of Workers	4
Images per Batch	2
Anchor Sizes	(8, 16, 32, 64, 128)
Anchor Aspect Ratios	(0.5, 1.0, 2.0)
RPN Batch Size per Image	512
RPN Positive Fraction	0.5
ROI Heads Batch Size per Image	1,024
NMS Testing Threshold	0.25
Max Iterations	5,0000
Learning Rate	0.001
Momentum	0.937
Weight Decay	0.0005
Test Detections per Image	256
Optimizer	SGD

IV. RESULTS AND DISCUSSIONS

The proposed method has a Mask R-CNN architecture with a *Backbone* ResNeXt-101-32 × 8d-FPN built using the open source library Detectron2 and pre-trained on the COCO dataset. To carry out the experiments, we used the method of dividing the dataset into training, validation, and testing in a proportion of 50:20:30, respectively. To perform the training and adjustment of the proposed method, the DDR dataset was used. The IDRiD dataset was also used to assess the generalizability of the proposed method. The evaluation was performed both in the detection (BBox) and in the segmentation (Mask) of fundus lesions being adopted the IoU according to Equation 1:

$$\text{IoU} = \frac{Overlap\ Area}{Union\ Area} \tag{1}$$

We compare the proposed method with different models that use state-of-the-art dense neural networks. The following models were used in the experiments: 1) Mask R-CNN ResNet-50 C4 [10], which uses a *Backbone* ResNet conv4 with a conv5 head [13]; 2) Mask R-CNN ResNet-50 C5-dilated [10], which uses *Backbone* ResNet conv5, with dilations in conv5 and standard heads Conv (convolutional layer) and FC (fully Connected Layer) for mask and bounding box prediction, respectively [27]; 3) Mask R-CNN ResNet-50 FPN [10], which uses a *Backbone* ResNet×FPN with standard Conv and FC heads for mask prediction and bounding box, respectively; 4) Mask R-CNN ResNet-101 C4 [10]; 5) Mask R-CNN ResNet-101 C5-dilated [10]; 6) Mask R-CNN ResNet-101 FPN [10]; and 6) Mask R-CNN ResNeXt-101-32 × 8d-FPN, as shown in Tables II and III. Table II presents the results obtained with the metric mAP for the limit of IoU of 0.5 with SGD optimizer. The proposed method using the *Backbone* ResNeXt-101-FPN achieved the best precision in the detection and segmentation tasks in the experiments performed in the validation set of the DDR dataset, as indicated in bold font, with a mAP of 0.2660.

TABLE II
RESULTS OBTAINED IN THE DETECTION AND SEGMENTATION TASKS OF FUNDUS LESIONS WITH THE METRIC mAP FOR THE LIMIT OF IoU OF 0.5 IN THE VALIDATION SET OF THE DDR DATASET WITH SGD OPTIMIZER.

Models		**Backbone**	*mAP*
BBox	Mask R-CNN	ResNet-50 C4	0.1325
	Mask R-CNN	ResNet-50 C5-dilated	0.1368
	Mask R-CNN	ResNet-50 FPN	0.1737
	Mask R-CNN	ResNet-101 C4	0.1746
	Mask R-CNN	ResNet-101 C5-dilated	0.1678
	Mask R-CNN	ResNet-101 FPN	0.1909
	Mask R-CNN	ResNeXt-101 FPN 32x8d	0.2124
	Proposed method	ResNeXt-101 FPN 32x8d	**0.2660**
Mask	Mask R-CNN	ResNet-50 C4	0.1617
	Mask R-CNN	ResNet-50 C5-dilated	0.1328
	Mask R-CNN	ResNet-50 FPN	0.1795
	Mask R-CNN	ResNet-101 C4	0.1624
	Mask R-CNN	ResNet-101 C5-dilated	0.1603
	Mask R-CNN	ResNet-101 FPN	0.1848
	Mask R-CNN	ResNeXt-101 FPN 32x8d	0.2205
	Proposed method	ResNeXt-101 FPN 32x8d	**0.2600**

We also evaluated and compared the results obtained by the proposed method on the IDRiD dataset, as presented in Table III. The proposed method using the *Backbone* ResNeXt-101-FPN obtained the best precision in the detection and segmentation tasks in the experiments carried out in the validation set of the IDRiD dataset using the SGD optimizer, as indicated in bold font, with mAP of 0.3460.

Fig. 3 presents the loss and Average Precision graphs obtained during the training and validation of the proposed method with SGD optimizer. In Fig. 3(d) are presented the results obtained with the detection of Bounding Boxes (BBox) of fundus lesions for the limit of IoU 0.50:0.95 in the validation using the DDR dataset. It is possible to verify that the proposed method reached greater precision in detecting Hard Exudates and less precision in detecting Microaneurysms.

To evaluate the accuracy of the proposed method we also calculated the AP and mAP with 10 IoU thresholds of IoU of 0.50:0.95 in the validation step of the DDR dataset for each lesion, as shown in Table IV. This metric rewards the detectors with the best location. The values of AP@[0.5:0.95] in the detection of retinal lesions were 0.0999, 0.1838, 0.0380, and

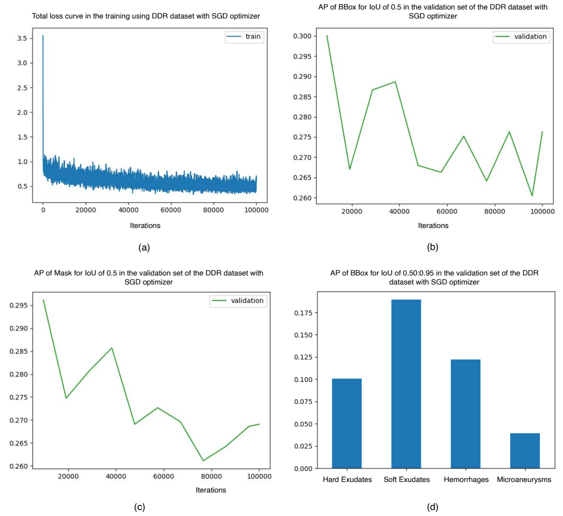

Fig. 3. Loss and Average Precision graphs obtained during training and validation of the proposed method. (a) Total loss curve in the training using DDR dataset with SGD optimizer. (b) AP of BBox for the limit of IoU of 0.5 in the validation set of the DDR dataset with SGD optimizer. (c) AP of Mask for the limit of IoU of 0.5 in the validation set of the DDR dataset with SGD optimizer. (d) AP of BBox for the limit of IoU of 0.50:0.95 in the validation set of the DDR dataset with SGD optimizer.

TABLE III
RESULTS OBTAINED IN THE DETECTION AND SEGMENTATION TASKS OF FUNDUS LESIONS WITH THE METRIC mAP FOR THE LIMIT OF IoU OF 0.5 IN THE VALIDATION SET OF THE IDRID DATASET WITH SGD OPTIMIZER.

	Models	Backbone	*mAP*
BBox	Mask R-CNN	ResNet-50 C4	0.2254
	Mask R-CNN	ResNet-50 C5-dilated	0.1789
	Mask R-CNN	ResNet-50 FPN	0.2050
	Mask R-CNN	ResNet-101 C4	0.2120
	Mask R-CNN	ResNet-101 C5-dilated	0.2065
	Mask R-CNN	ResNet-101 FPN	0.1656
	Mask R-CNN	ResNeXt-101 FPN 32x8d	0.2188
	Proposed method	ResNeXt-101 FPN 32x8d	**0.3460**
Mask	Mask R-CNN	ResNet-50 C4	0.1921
	Mask R-CNN	ResNet-50 C5-dilated	0.1645
	Mask R-CNN	ResNet-50 FPN	0.2138
	Mask R-CNN	ResNet-101 C4	0.2161
	Mask R-CNN	ResNet-101 C5-dilated	0.1841
	Mask R-CNN	ResNet-101 FPN	0.1789
	Mask R-CNN	ResNeXt-101 FPN 32x8d	0.2050
	Proposed method	ResNeXt-101 FPN 32x8d	**0.3210**

0.1183, for EX, SE, MA and HE, respectively, as shown in Table IV. To better understand the results obtained, we present in Fig. 4 an example of instance segmentation performed by the proposed method on a fundus image of the DDR dataset. We used a `NMS Testing Threshold` value of 0.25. It is possible to verify that the locations *pixel* to *pixel* of the detected lesions were obtained, as well as the bounding boxes of each lesion. This type of approach can more effectively support the medical diagnosis, since it is possible to visualize with greater clarity the extension and size of the lesion, as opposed to just detecting and tracing a bounding box around the lesion in the image.

TABLE IV
RESULTS OBTAINED WITH DETECTION (BBOX) AND SEGMENTATION (MASK) BY THE PROPOSED METHOD WITH AP AND mAP WITH 10 IoU THRESHOLDS OF 0.50:0.95 IN THE VALIDATION SET OF THE DDR DATASET.

Model		*AP*				*mAP*
		EX	**SE**	**MA**	**HE**	
Proposed method	BBox	0.0999	0.1838	0.0380	0.1183	0.1100
	Mask	0.0994	0.1893	0.0412	0.1072	0.1093

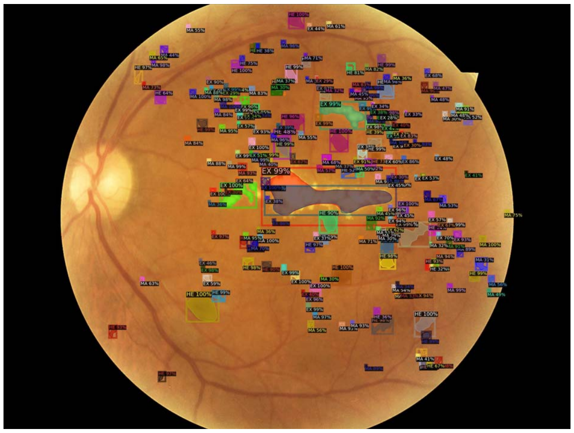

Fig. 4. Instance segmentation of fundus lesions performed by the proposed method in the fundus image "007-3892-200.jpg" of the test set of the DDR dataset. The classification of the lesions identified in the image was performed at the *pixel* level, with the lesion label and the percentage of confidence associated with the detected object being assigned. Each instance segmentation has a different color regardless of the lesion class.

It is verified that the proposed method that the proposed method obtained promising results in the instance segmentation of the investigated fundus lesions, even with the presence of multiple lesions with variable sizes and shapes. Because the Mask R-CNN model has an RPN module, it is possible to extract ROIs from the image that is more likely to contain fundus lesions. Experimental results showed that the proposed method presents greater precision in the detection of soft exudates and less precision in the detection of Microaneurysms. The difficulty in detecting microaneurysms is due to the small size of this type of lesion. During model training, there is a more accentuated gradient dissipation in the extraction of features from very small objects, which ends up causing a deficient precision of these objects due to the neural network confusing the microlesion with the image background, generating in turn high rates of false negatives, for example.

V. CONCLUSIONS

This article presented a two-stage detector-based method for instance segmentation of DR-associated lesions. The neural network architecture was built based on the Mask R-CNN model using the open-source library Detectron2. For training and evaluation of the proposed method, public datasets of

DDR and IDRiD diabetic retinopathy were used. The datasets were divided into training, validation, and test sets in a ratio of 50:20:30, respectively. With the IoU threshold of 0.5, the lesion instance targeting reached mAP of 0.2660 in the validation step and mAP of 0.1600 in the test step.

The proposed method obtained AP@[0.5:0.95] of 0.0999 for Hard Exudates; 0.1838 for Soft Exudates; 0.0380 for Microaneurysms; and 0.1183 for Hemorrhages. The results obtained were promising, demonstrating that the instance segmentation of fundus lesions performed through deep neural networks can help in medical diagnosis. However, the results presented in this work indicate that the segmentation of fundus lesions is extremely difficult and represents a challenge for future research.

In future work, we intend to develop solutions that combine different contexts, such as the classification, detection, and segmentation of lesions associated with DR to obtain more accurate results. We also aim to improve the pre-processing and data augmentation methods of retinal images, to provide a more efficient feature extraction from microlesions.

Acknowledgment

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

[1] P. Amerikanos and I. Maglogiannis, "Image analysis in digital pathology utilizing machine learning and deep neural networks," *Journal of Personalized Medicine*, vol. 12, no. 9, 2022. [Online]. Available: https://www.mdpi.com/2075-4426/12/9/1444

[2] P. Krishnadas, K. Chadaga, N. Sampathila, S. Rao, S. K. S., and S. Prabhu, "Classification of malaria using object detection models," *Informatics*, vol. 9, no. 4, 2022. [Online]. Available: https://www.mdpi.com/2227-9709/9/4/76

[3] International Council of Ophthalmology, "Updated 2017 ICO Guidelines for Diabetic Eye Care," *ICO Guidelines for Diabetic Eye Care*, pp. 1–33, 2017. [Online]. Available: http://www.icoph.org/downloads/ICOGuidelinesforDiabeticEyeCare.pdf

[4] D. S. W. Ting, G. C. M. Cheung, and T. Y. Wong, "Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review," *Clinical and Experimental Ophthalmology*, vol. 44, no. 4, pp. 260–277, 2016.

[5] E. Vocaturo and E. Zumpano, "The contribution of AI in the detection of the Diabetic Retinopathy," *Proceedings - 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020*, pp. 1516–1519, 2020.

[6] D. X. Fei-Fei Li, Ranjay Krishna, "cs231n, Lecture 15 - Slide 4, Detection and Segmentation," http://cs231n.stanford.edu/slides/2021/lecture_15.pdf, 2021, [Online; accessed 26-December-2021].

[7] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, and H. Kang, "Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening," *Information Sciences*, vol. 501, pp. 511–522, 2019. [Online]. Available: https://doi.org/10.1016/j.ins.2019.06.011

[8] L. Dai, L. Wu, H. Li, C. Cai, Q. Wu, H. Kong, R. Liu, X. Wang, X. Hou, Y. Liu, X. Long, Y. Wen, L. Lu, Y. Shen, Y. Chen, D. Shen, X. Yang, H. Zou, B. Sheng, and W. Jia, "A deep learning system for detecting diabetic retinopathy across the disease spectrum," *Nature Communications*, vol. 12, no. 1, 2021. [Online]. Available: http://dx.doi.org/10.1038/s41467-021-23458-5

[9] F. Shenavarmasouleh, F. G. Mohammadi, M. H. Amini, T. Taha, K. Rasheed, and H. R. Arabnia, "Drdrv3: Complete lesion detection in fundus images using mask r-cnn, transfer learning, and lstm," 2021. [Online]. Available: https://arxiv.org/abs/2108.08095

[10] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 386–397, 2020.

[11] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[12] A. Hong, G. Lee, H. Lee, J. Seo, and D. Yeo, "Deep learning model generalization with ensemble in endoscopic images," *CEUR Workshop Proceedings*, vol. 2886, pp. 80–89, 2021.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[14] Z.-Q. Zhao, P. Zheng, S. tao Xu, and X. Wu, "Object detection with deep learning: A review," 2019.

[15] Y. Konishi, Y. Hanzawa, M. Kawade, and M. Hashimoto, "SSD: Single Shot MultiBox Detector," *Eccv*, vol. 1, pp. 398–413, 2016.

[16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, Las Vegas, NV, USA, 27–30 June 2016, 2016, pp. 779–788.

[17] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu, "A survey of deep learning-based object detection," *IEEE Access*, vol. 7, pp. 128 837–128 868, 2019.

[18] J. A. Kim, J. Y. Sung, and S. H. Park, "Comparison of Faster-RCNN, YOLO, and SSD for Real-Time Vehicle Type Recognition," *2020 IEEE International Conference on Consumer Electronics - Asia, ICCE-Asia 2020*, pp. 8–11, 2020.

[19] C. B. Murthy, M. F. Hashmi, N. D. Bokde, and Z. W. Geem, "Investigations of object detection in images/videos using various deep learning techniques and embedded platforms—a comprehensive review," *Applied Sciences*, vol. 10, no. 9, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/9/3280

[20] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, and F. Meriaudeau, "Indian diabetic retinopathy image dataset (idrid): A database for diabetic retinopathy screening research," *Data*, vol. 3, no. 3, 2018. [Online]. Available: https://www.mdpi.com/2306-5729/3/3/25

[21] C. Santos, M. Aguiar, D. Welfer, and B. Belloni, "A new approach for detecting fundus lesions using image processing and deep neural network architecture based on yolo model," *Sensors*, vol. 22, no. 17, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/17/6441

[22] W. L. Alyoubi, M. F. Abulkhair, and W. M. Shalash, "Diabetic retinopathy fundus image classification and lesions localization system using deep learning," *Sensors*, vol. 21, no. 11, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/11/3704

[23] M. S. Nixon and A. S. Aguado, "5 - high-level feature extraction: fixed shape matching," in *Feature Extraction and Image Processing for Computer Vision (Fourth Edition)*, 4th ed., M. S. Nixon and A. S. Aguado, Eds. Academic Press, 2020, pp. 223–290. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128149768000051

[24] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 5987–5995, 2017.

[25] X. Li, T. Lai, S. Wang, Q. Chen, C. Yang, and R. Chen, "Feature Pyramid Networks for Object Detection," *Proceedings - 2019 IEEE Intl Conf on Parallel and Distributed Processing with Applications, Big Data and Cloud Computing, Sustainable Computing and Communications, Social Computing and Networking, ISPA/BDCloud/SustainCom/SocialCom 2019*, pp. 1500–1504, 2019.

[26] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8693 LNCS, no. PART 5, pp. 740–755, 2014.

[27] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable Convolutional Networks," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-Octob, pp. 764–773, 2017.