Artigo Original
DOI: https://doi.org/10.59681/2175-4411.v18.2026.1510

JHI JOURNAL OF HEALTH INFORMATICS

BDR-iD: A Preliminary Brazilian Dataset of Retinal Lesions and Deep Learning Baselines for Diabetic Retinopathy

BDR-iD: Conjunto de Dados Brasileiro Preliminar de Lesões Retinianas e Linhas de Base em Aprendizado Profundo para Retinopatia Diabética

BDR-iD: Conjunto de datos brasileño preliminar de lesiones retinianas y líneas base de aprendizaje profundo para retinopatía diabética

Carlos Santos[1], Laura Bernardes[2], Artur Heckler[3], Alejandro Pereira[4], Marcelo Dias[5], Marilton Aguiar[6], Daniel Welfer[7]

ABSTRACT

Objective: To present the construction and initial characterization of the Brazilian BDR-iD dataset of fundus images for diabetic retinopathy (DR) research, and to define deep learning baselines for DR grading, lesion segmentation, and lesion detection. DR is a major cause of visual impairment in adults, and early diagnosis is critical; however, limited infrastructure and specialist availability restrict access, particularly in settings comparable to Brazil's public health system. We collected and anonymized 13,131 fundus images from a clinic in Pelotas, Brazil, acquired between 2012 and 2024. From this cohort, 150 images were selected and expert-annotated for DR presence and lesion findings, including microaneurysms, hemorrhages, and exudates. Models were evaluated on three tasks: DR severity grading, lesion segmentation, and lesion detection. For severity classification, the best test-set baseline achieved an overall accuracy (OA) of 0.6667. Segmentation and detection showed more modest performance, reflecting the limited number of annotated images, class imbalance, and the intrinsic difficulty of microlesion annotation. The preliminary BDR-iD release is not intended for clinical deployment, but as a starting point toward larger and more standardized Brazilian datasets, providing a public resource and reference baselines for future national studies.

Keywords: Diabetic retinopathy, classification, segmentation, deep learning

RESUMO

Objetivo: Apresentar a construção e a caracterização inicial do conjunto de dados brasileiro BDR-iD, com imagens de fundo para estudo da retinopatia diabética (RD), e definir linhas de base de aprendizado profundo para graduação da RD, segmentação e detecção de lesões. A RD é uma causa relevante de perda visual em adultos e o diagnóstico precoce é crítico; porém, a falta de infraestrutura e de especialistas limita o acesso, especialmente em contextos semelhantes ao SUS. Foram coletadas e anonimizadas 13.131 imagens de uma clínica em Pelotas (Brasil), entre 2012 e 2024. Dentre elas, 150 foram selecionadas e anotadas por especialista quanto à presença de RD e de lesões, incluindo microaneurismas, hemorragias e exsudatos. Modelos foram avaliados em três tarefas: graduação da RD, segmentação de lesões e detecção de lesões. Na classificação de severidade, no conjunto de teste, a melhor linha de base atingiu acurácia global (OA) de 0,6667. Segmentação e detecção apresentaram desempenho mais modesto, refletindo poucas imagens anotadas, desbalanceamento de classes e a dificuldade intrínseca de anotar microlesões. O BDR-iD preliminar não deve ser usado para implantação clínica, mas como ponto de partida para bases brasileiras mais amplas e padronizadas, oferecendo um recurso público inicial e referenciais para estudos futuros.

Descritores: Retinopatia diabética, classificação, segmentação, redes neurais profundas

RESUMEN

Objetivo: Presentamos la construcción y caracterización inicial del conjunto de datos brasileño BDR-iD, con imágenes de fondo de ojo para estudiar la retinopatía diabética (RD), y establecemos líneas base de aprendizaje profundo para graduación, segmentación y detección de lesiones. La RD causa pérdida visual en adultos y el diagnóstico temprano es clave; sin embargo, la falta de infraestructura y especialistas limita el acceso, especialmente en entornos similares al sistema público brasileño. Se recopilaron y anonimizaron 13.131 imágenes de una clínica en Pelotas (Brasil) entre 2012 y 2024. De ellas, 150 fueron seleccionadas y anotadas por un especialista para presencia de RD y lesiones (microaneurismas, hemorragias y exudados). Los modelos se evaluaron en tres tareas. En severidad, la mejor línea base en el conjunto de prueba logró OA=0,6667. Segmentación y detección mostraron rendimiento más modesto por el bajo número de anotaciones, el desbalance de clases y la dificultad de anotar microlesiones. Esta versión preliminar no es para uso clínico, sino un punto de partida hacia conjuntos brasileños más amplios y estandarizados, con un recurso público y baselines para estudios futuros.

Descriptores: Diabetic retinopathy, classification, segmentation, deep learning

[1] *Doutor em Ciência da Computação, Instituto Federal Farroupilha (IFFar).*
[2] *Graduanda em Tecnologia em Análise e Desenvolvimento de Sistemas, Instituto Federal Farroupilha (IFFar).*
[3] *Estudante do Curso Técnico em Informática Integrado ao Ensino Médio, Instituto Federal Farroupilha (IFFar).*
[4] *Mestre em Ciência da Computação, Universidade Federal de Pelotas (UFPel).*
[5] *Doutor em Ciência da Computação, Universidade Federal de Pelotas (UFPel).*
[6] *Doutor em Ciência da Computação, Universidade Federal de Santa Maria (UFSM).*

Autor Correspondente: **Carlos Santos**
e-mail: **carlos.santos@iffarroupilha.edu.br**

Artigo recebido: 08/06/2025
Aprovado: 07/01/2026
https://jhi.sbis.org.br/

INTRODUCTION

According to the International Council of Ophthalmology[1], Diabetic Retinopathy (DR) is a major cause of vision loss in working-age adults, affecting approximately one-third (34.6%) of people with diabetes in the US, Europe, and Asia. The increasing global prevalence of diabetes is expected to raise vision loss due to related complications. DR is a leading cause of visual impairment among individuals aged 20–74, and its diagnosis relies on identifying retinal lesions, such as microaneurysms (MA), hemorrhages (HE), soft exudates (SE), and hard exudates (EX)[2].

In Brazil, organizing DR screening and follow-up is particularly challenging because of the country's large territory, uneven distribution of ophthalmologists, and high demand on the public health system (Sistema Único de Saúde – SUS). These constraints contribute to delays in diagnosis and treatment. Computer-aided diagnosis and deep learning (DL) models are therefore promising for supporting ophthalmologists and primary-care professionals, enabling more scalable screening and helping prioritize higher-risk patients.

Medical image analysis is central to early DR detection: timely treatment can prevent vision loss[3]. However, limited examination capacity and a global shortage of ophthalmologists — especially in developing regions — remain key barriers[4]. Although DL has shown strong potential for lesion detection and segmentation, progress is constrained by the scarcity of publicly available, expert-annotated fundus datasets. In Brazil, the lack of high-quality, nationally representative datasets further limits the development and validation of reliable DL systems. Building such datasets requires ethical approval, rigorous anonymization, and substantial expert annotation effort, but it is essential for transparency, reproducibility, and trust in AI-assisted medical decision-making.

In this context, this study aims to: (i) assemble and anonymize a preliminary Brazilian fundus dataset for DR analysis (BDR-iD); (ii) obtain expert lesion-level annotations for a subset spanning different DR stages; and (iii) benchmark state-of-the-art DL models for DR grading, lesion segmentation, and lesion detection. Rather than presenting a definitive clinical tool, we introduce BDR-iD as an initial public resource and provide baseline results to guide future DR research in Brazil.

METHODOLOGY

This work introduces the Brazilian Diabetic Retinopathy Images Dataset (BDR-iD), built from 13,131 anonymized fundus images collected at an ophthalmology clinic in Pelotas, Brazil (2012–2024) using a 45° field-of-view (FOV) Canon CX1 retinograph with a Canon BM7-0331 camera. Patients ranged from 0 to 99 years (mean age: 58 years), with 58.82% female and 41.18% male. Because lesion annotation is complex, expert labeling was limited to a curated subset: 150 images for DR grading/classification and 100 images for lesion detection and segmentation. The subset was obtained through quality filtering to remove low-resolution, dark, glare-affected, and blurred images, followed by manual exclusions. In total, 150 images were classified and/or annotated for DR grading, lesion semantic segmentation, and lesion detection.

Retinal lesions for image detection and segmentation include EX, SE, MA, and HE. The segmentation process uses 100 images, of which 38 show DR and were annotated at the pixel level. Table 1 lists the annotations for each lesion type in the BDR-iD dataset. The dataset includes three annotation types: image-level DR classification, pixel-level masks, and bounding boxes for lesions. Annotations were initially generated automatically and later validated by a medical expert. In this version, the EX and SE annotations were fully validated.

Table 1 - Distribution of annotated fundus lesions in the BDR-iD dataset for segmentation and detection tasks.

Lesions	Quantity
hard exudates	974
hemorrhages	318
soft exudates	17
microaneurysms	307
Total	1616

Source: Prepared by the authors.

All annotated lesions were included to enable a comprehensive evaluation of segmentation and detection. A specialist classified DR based on medical reports, yielding 88 DR cases, 54 healthy images, and eight images with an undefined DR stage. Lesion masks were initially generated by R2U-Net at 256×256×3 and then upscaled to the original resolution using Upscayl/Real-ESRGAN; edges were converted to polygons and exported as editable COCO annotations. These automatic labels were then manually validated and corrected in CVAT by a retina specialist, who also added missed lesions. Microaneurysms require manual annotation because of their very small size and may even require fluorescein angiography for reliable detection.

After validation, the dataset was finalized as a COCO instance-segmentation file and corresponding binary masks for semantic segmentation, including bounding boxes and updated polygons. Building BDR-iD v1 was time-consuming and heavily dependent on expert validation, particularly for small lesions such as microaneurysms and hemorrhages (14–136 μm). As a result, v1 remains limited in the number of images and small-lesion annotations; future releases will expand the dataset and improve segmentation and detection coverage. All v1 annotations were produced by a single specialist, and future versions will include multiple graders to report inter-observer agreement (e.g., Kappa).

RESULTS AND DISCUSSION

The evaluation of BDR-iD considered three complementary computer vision tasks: DR grading, lesion segmentation, and lesion detection. Together, these experiments provide a baseline characterization of the performance of current deep learning models on this preliminary Brazilian dataset and highlight the main challenges for future work.

Deep learning models for DR grading

The dataset was randomly split into training (50%), validation (20%), and test (30%) sets (class distribution in Table 2). Training images were resized to 512×512 and augmented with rotations (≤30°), horizontal/vertical flips, ColorJitter (≤20% for brightness/contrast/saturation/hue), and random cropping. Class imbalance was mitigated with a batch size of 16 and a Weighted Random Sampler. Models were trained for 200 epochs using Adam (lr=0.001).

Table 2 - Class distribution across the train, validation, and test sets in the DR classification task.

Split	No DR	Mild NPDR	Moderate NPDR	Severe NPDR	PDR	Unclassifiable
Train	25	10	14	8	13	5
Validation	8	5	6	4	6	1
Test	21	8	6	3	5	2

Source: Prepared by the authors.

DR classification results are reported in Table 3(a) using per-class accuracy (No DR, Mild/Moderate/Severe NPDR, PDR, Unclassifiable), Overall Accuracy (OA), Average Accuracy (AA), and Kappa. VGG-16(5), SwinTransformer, and SqueezeNet performed poorly: VGG-16 reached 27% accuracy for No DR with OA 26.67%, AA 4.44%, and Kappa 0.0000 (near-random). SwinTransformer achieved 47% for No DR and 33% for PDR (AA 13.40%, Kappa 0.19). SqueezeNet reached 44% for No DR and 50% for Severe NPDR, with OA 40%.

Table 3 - Results obtained in the classification for (a) the validation set and (b) the test set.

Models	**No DR**	**Mild NPDR**	**Moderate NPDR**	**Severe NPDR**	**PDR**	**Unclassifiable**	**OA**	**AA**	**Kappa**
VGG-16	0.2700	0.0000	0.0000	0.0000	0.0000	0.0000	0.2667	0.0444	0.0000
ResNet-18	0.8600	0.6700	0.6700	0.4000	0.6200	0.0000	0.6333	0.5359	0.5461
GoogLeNet	0.8000	1.0000	0.6700	0.7500	0.5700	1.0000	0.7333	0.7980	0.6643
DenseNet-121	0.7300	0.0000	0.5000	0.3800	0.8000	0.0000	0.6000	0.4004	0.4958
EfficientNet B0	1.0000	1.0000	0.6700	0.0000	0.6200	0.0000	0.8000	0.5486	0.7461
RegNet Y 400MF	1.0000	1.0000	0.5000	1.0000	0.6000	0.0000	0.7333	0.6833	0.6667
SqueezeNet	0.4400	0.0000	0.0000	0.5000	0.3000	0.0000	0.4000	0.2074	0.2151
SwinTransformer	0.4700	0.0000	0.0000	0.0000	0.3300	0.0000	0.3667	0.1340	0.1926

(a)

Models	No DR	Mild NPDR	Moderate NPDR	Severe NPDR	PDR	Unclassifiable	OA	AA	Kappa
VGG-16	0.4700	0.0000	0.0000	0.0000	0.0000	0.0000	0.4667	0.0778	0.0000
ResNet-18	0.7200	0.5000	0.5000	0.3300	0.4300	0.0000	0.6222	0.4143	0.4239
GoogLeNet	0.7800	0.5000	0.0000	0.1700	0.4000	0.0000	0.6000	0.3074	0.4048
DenseNet-121	0.7300	1.0000	0.3300	0.1200	0.4000	0.0000	0.4889	0.4309	0.2930
EfficientNet B0	0.7900	0.5700	0.6000	0.3300	0.4000	0.0000	0.6444	0.4494	0.4853
RegNet Y 400MF	0.8300	0.4500	0.3300	0.5000	0.6700	0.0000	0.6667	0.4634	0.5179
SqueezeNet	0.6000	0.0000	0.0000	0.0000	0.2100	0.0000	0.3556	0.1351	0.1265
SwinTransformer	0.5000	0.0000	0.0000	0.0000	0.1700	0.0000	0.2889	0.1111	0.0400

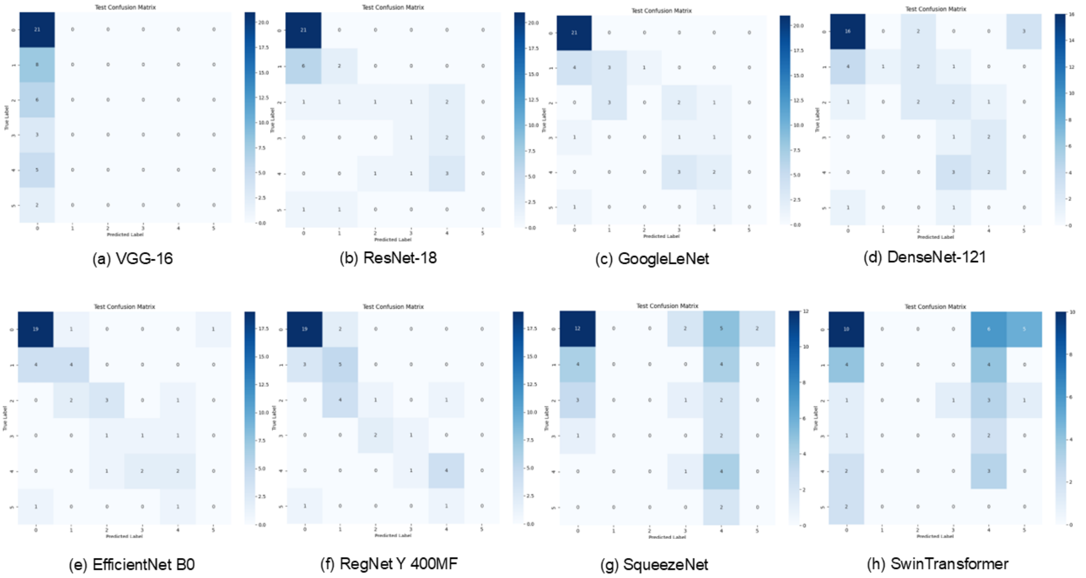
(b)

Source: Prepared by the authors.

DenseNet-121 and ResNet-18 achieved median results: DenseNet-121 scored 73% accuracy for No DR and 80% for PDR, but struggled with Mild NPDR, achieving low accuracy for Moderate (50%) and Severe NPDR (38%). Its OA was 60%, AA was 40%, and Kappa was 0.49, indicating moderate but inconsistent performance. ResNet-18 performed well in most classes, especially No DR (86%), Mild NPDR (67%), Moderate NPDR (67%), and PDR (62%), but poorly on Severe NPDR (40%), with OA 63.33%, AA 53.59%, and Kappa 0.54, indicating a more reliable model that needs further tuning.

Figure 1 shows the test-set confusion matrices for the evaluated architectures (VGG-16, ResNet-18(6), GoogLeNet(7), DenseNet-121(8), EfficientNet B0(9), RegNet Y 400MF(10), SqueezeNet(11), SwinTransfor-

Figure 1 - Confusion matrices of the experiments performed on the test set of the BDR-iD dataset.



(a) VGG-16 (b) ResNet-18 (c) GoogleLeNet (d) DenseNet-121

(e) EfficientNet B0 (f) RegNet Y 400MF (g) SqueezeNet (h) SwinTransformer

Source: Prepared by the authors.

The top classifiers were EfficientNet-B0, RegNetY-400MF, and GoogLeNet. On the validation set, EfficientNet-B0 achieved 100% accuracy for No DR and Mild NPDR, moderate performance for Moderate NPDR (67%) and PDR (62%), but 0% for Severe NPDR, yielding OA = 0.80, AA = 54.86%, and Kappa = 0.74 (highest Kappa). RegNetY-400MF reached 100% for No DR, Mild NPDR, and Severe NPDR, with 50% for Moderate NPDR and 60% for PDR (OA = 0.7333, AA = 68.33%, Kappa = 0.67). GoogLeNet was more consistent across classes (OA = 0.7333, AA = 79.80%, Kappa = 0.66). Overall, GoogLeNet showed the most consistent class-wise performance, whereas RegNetY-400MF was more balanced in terms of per-class accuracy distribution (Table 3(b)).

Despite these highlights, models generally underperformed on the test set, suggesting overfitting and limited generalization, especially for classes with few samples. In the final test-set comparison, RegNet Y 400MF and EfficientNet B0 remained the top performers, with RegNet Y 400MF leading (OA 0.6667, AA 0.4634, Kappa 0.5179) and showing relative robustness in PDR (0.67) and Severe NPDR (0.50), while EfficientNet B0 achieved OA 0.6444 and Kappa 0.4853. The results underscore the need for more data and improved preprocessing, particularly for intermediate classes.

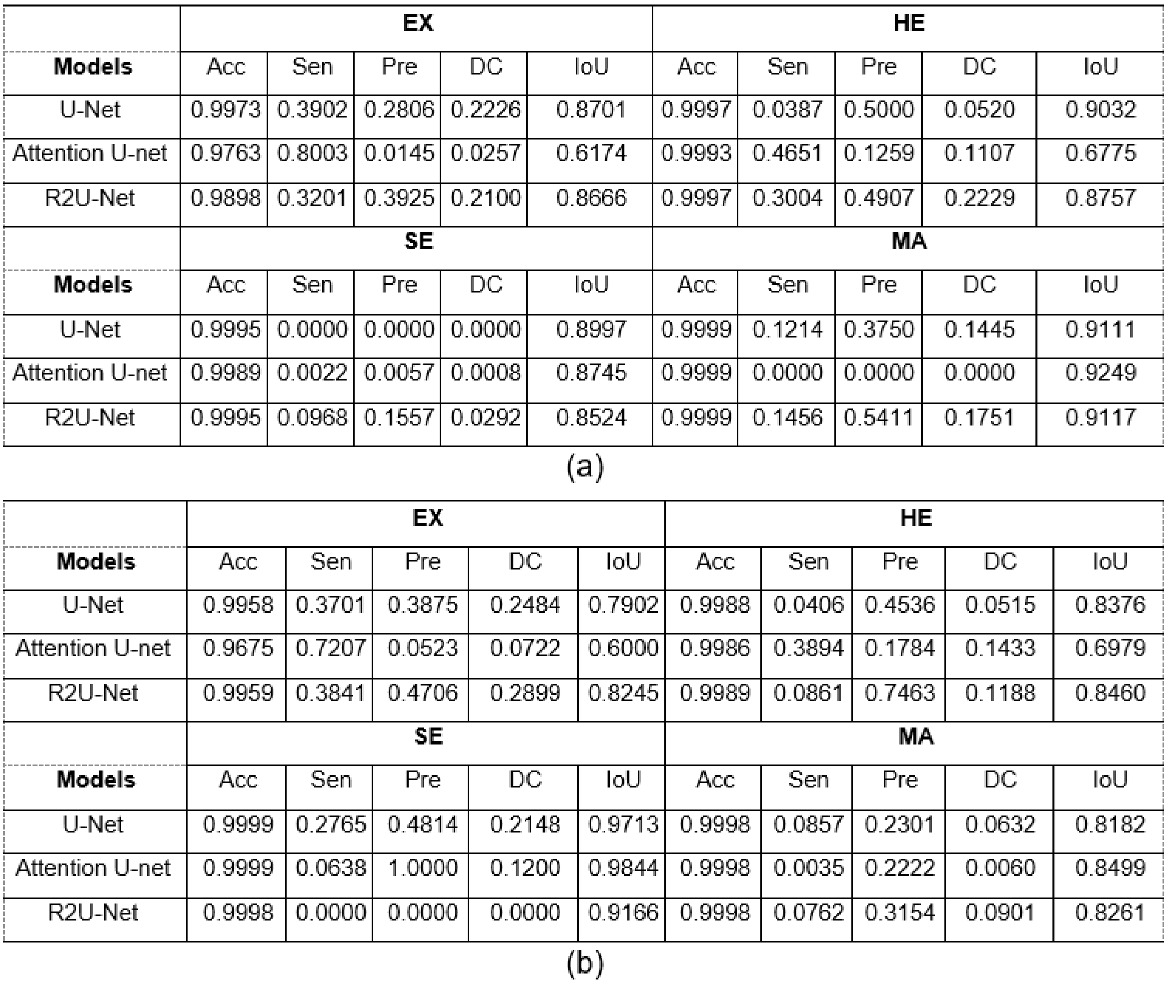
Deep learning models for lesion segmentation

Models were trained for lesion segmentation with a batch size of 4 over 50 epochs on 256×256 inputs, using Adam (lr=0.001), ReLU in hidden layers, Sigmoid output, and ImageNet initialization. Data augmentation included horizontal/vertical flips, elastic transform, grid distortion, and optical distortion, generating five augmented variants per training image to increase diversity and reduce overfitting.

Segmentation was evaluated on BDR-iD using pixel accuracy (Acc), sensitivity (recall), precision, Dice, and IoU, comparing U-Net[13], Attention U-Net[14], and R2U-Net[15] across hard exudates, hemorrhages, soft exudates, and microaneurysms (Table 4(a)). Although overall accuracy was high, it largely reflected background dominance in this highly imbalanced setting; therefore, Dice, sensitivity, and precision provide a more informative assessment of lesion overlap and detection behavior. Sensitivity varied substantially across models and lesions. Soft exudates and microaneurysms exhibited very low sensitivity, indicating frequent detection failures. For hard exudates, Attention U-Net achieved higher sensitivity but lower precision (more false positives), whereas U-Net and R2U-Net provided a better sensitivity–precision trade-off. For hemorrhages, R2U-Net achieved the best Dice, but with low sensitivity.

Given the strong class imbalance, many images contain no pixels for a given lesion class. In such cases, IoU may become undefined (union=0) and can be affected by the adopted convention. Therefore, we emphasize Dice/recall/precision and recommend reporting IoU/Dice conditioned on positive ground-truth cases.

Table 4 - Results obtained from the segmentation for (a) the validation set and (b) the test set.



	EX					HE				
Models	Acc	Sen	Pre	DC	IoU	Acc	Sen	Pre	DC	IoU
U-Net	0.9973	0.3902	0.2806	0.2226	0.8701	0.9997	0.0387	0.5000	0.0520	0.9032
Attention U-net	0.9763	0.8003	0.0145	0.0257	0.6174	0.9993	0.4651	0.1259	0.1107	0.6775
R2U-Net	0.9898	0.3201	0.3925	0.2100	0.8666	0.9997	0.3004	0.4907	0.2229	0.8757
	SE					**MA**				
Models	Acc	Sen	Pre	DC	IoU	Acc	Sen	Pre	DC	IoU
U-Net	0.9995	0.0000	0.0000	0.0000	0.8997	0.9999	0.1214	0.3750	0.1445	0.9111
Attention U-net	0.9989	0.0022	0.0057	0.0008	0.8745	0.9999	0.0000	0.0000	0.0000	0.9249
R2U-Net	0.9995	0.0968	0.1557	0.0292	0.8524	0.9999	0.1456	0.5411	0.1751	0.9117

(a)

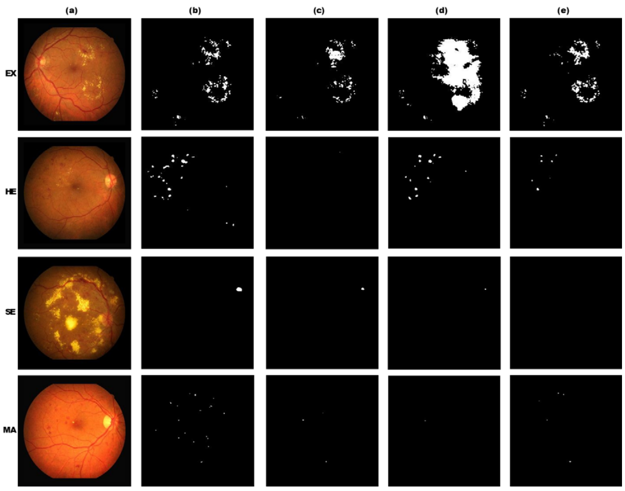
	EX					HE				
Models	Acc	Sen	Pre	DC	IoU	Acc	Sen	Pre	DC	IoU
U-Net	0.9958	0.3701	0.3875	0.2484	0.7902	0.9988	0.0406	0.4536	0.0515	0.8376
Attention U-net	0.9675	0.7207	0.0523	0.0722	0.6000	0.9986	0.3894	0.1784	0.1433	0.6979
R2U-Net	0.9959	0.3841	0.4706	0.2899	0.8245	0.9989	0.0861	0.7463	0.1188	0.8460
	SE					**MA**				
Models	Acc	Sen	Pre	DC	IoU	Acc	Sen	Pre	DC	IoU
U-Net	0.9999	0.2765	0.4814	0.2148	0.9713	0.9998	0.0857	0.2301	0.0632	0.8182
Attention U-net	0.9999	0.0638	1.0000	0.1200	0.9844	0.9998	0.0035	0.2222	0.0060	0.8499
R2U-Net	0.9998	0.0000	0.0000	0.0000	0.9166	0.9998	0.0762	0.3154	0.0901	0.8261

(b)

Source: Prepared by the authors.

On the test set, accuracy remained high, yet sensitivity continued to depend on lesion type. R2U-Net delivered the best precision and Dice for hard exudates; Attention U-Net found more lesions but produced many false positives. For hemorrhages, R2U-Net showed high precision but low sensitivity, while Attention U-Net exhibited the opposite. For soft exudates and microaneurysms, all models maintained low sensitivity, underscoring the difficulty of reliable segmentation with limited annotations.

Figure 2 - Comparison between the fundus lesion segmentations performed by the models with ground truth in images from the test set of the BDR-iD dataset.



Source: Prepared by the authors.

Figure 2 presents a visual comparison of fundus lesion segmentation results produced by different models on test-set images from the BDR-iD dataset. Figure 2(a) shows the original fundus image, while Figure 2(b) shows the ground-truth lesion mask. Figures 2(c), (d), and (e) display the predictions of U-Net, Attention U-Net, and R2U-Net, respectively.

Deep learning models for lesion detection

Models were trained for lesion detection with batch size 8 over 50 epochs on 640×640 images, using early stopping (patience 15) and Rectified Adam (lr=0.001) with pre-trained YOLOv9, YOLOv10, and YOLOv11. Data augmentation (flips, distortions, brightness changes) produced five variants per training image to reduce overfitting. Performance was measured using mAP@50 across four lesion classes (EX, HE, SE, MA).

YOLOv9 achieved intermediate performance (mAP 0.3170), performing better on HE and SE but weaker on EX and MA. YOLOv10 was the weakest (mAP 0.1880) and failed to detect SE. YOLOv11 performed best (mAP 0.3460), excelling on SE (0.6670) and showing the most balanced results, though MA remained low. During YOLOv11 training, losses decreased and precision/recall improved, while mAP plateaued around 0.23, suggesting a reasonable fit with room to improve. On the test set, YOLOv9's mAP slightly decreased (-0.0210): EX detection improved, but HE and MA worsened, consistent with possible overfitting; SE performance remained stable.

Table 5 - Results obtained in lesion detection when compared to models using the mAP@50 metrics in (a) the validation set and (b) the test set.

Models	Yolov9	Yolov10	YOLOv11
EX	0.0914	0.0967	0.1170
HE	0.3990	0.3980	0.3790
SE	0.5010	0.0000	0.6670
MA	0.2780	0.2580	0.2220
mAP	0.3170	0.1880	0.3460

(a)

Models	Yolov9	Yolov10	YOLOv11
EX	0.2480	0.1530	0.2210
HE	0.2460	0.3640	0.3560
SE	0.5340	0.0000	0.7500
MA	0.1560	0.2290	0.1980
mAP	0.2960	0.1870	0.3810

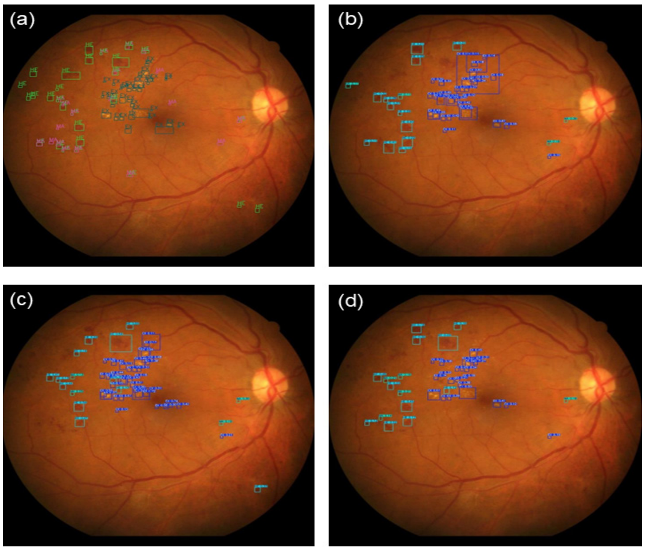
(b)

Source: Prepared by the authors.

The YOLOv10 model showed similar performance in validation and testing, indicating potential underfitting due to consistently low results. The primary concern is SE (0.0000), indicating that the model did not learn to detect this class. The detection of HE (0.3640) and MA (0.2290) improved, yet remained below that of other models. The decrease in EX detection (0.1530) suggests minimal learning during testing. The YOLOv11 model showed improvements on the test set, indicating enhanced generalization ability. Its mAP increased from 0.3460 to 0.3810 (Table 5(a) and (b), respectively), boosting its performance in detecting SE (0.7500) and refining its representation of this class. However, it experienced a decrease in MA detection (0.1980), highlighting the model's overall challenges.

Figure 3 illustrates lesion detection examples produced by YOLOv9, YOLOv10, and YOLOv11 on a BDR-iD fundus image. Although microlesion detection remains challenging, the models identified a substantial number of lesions, indicating that BDR-iD can support training state-of-the-art deep learning systems for improved medical diagnosis.

Figure 3 - Fundus lesion detection performed by YOLOv9, YOLOv10, and YOLOv11 models on an image from the test set of the BDR-iD dataset.



Source: Prepared by the authors.

Across tasks, classification showed overfitting and limited generalization, particularly when separating similar classes such as Moderate vs. Severe NPDR under data scarcity. In segmentation, a sensitivity–accuracy trade-off emerged: Attention U-Net achieved higher sensitivity but generated many false positives, while R2U-Net achieved higher accuracy but missed lesions. Soft exudates (SE) and microaneurysms (MA) consistently underperformed due to low prevalence and annotation difficulty, and YOLO detectors faced similar constraints—especially YOLOv9, which appeared more prone to overfitting—whereas YOLOv11 generalized better. The proposed next steps are to expand the labeled set, improve effective image resolution, and fine-tune models to mitigate these limitations.

These results should be interpreted as baseline evidence, not as performance suitable for clinical deployment. Large-scale DR screening studies trained on extensive datasets report markedly higher performance. For instance, Gulshan et al. trained on over 120,000 EyePACS and Messidor-2 images and achieved sensitivities up to 97.5% with specificities above 93% for referable DR in external validation(16). Ting et al. similarly validated a system in multiethnic diabetic populations, reporting 90.5% sensitivity and 91.6% specificity for referable DR(17). These benchmarks contextualize the more modest results from BDR-iD v1, which are expected due to the small annotated subset, single-center data, and class imbalance (notably for microaneurysms and soft exudates). Moreover, lesion-level annotation is inherently difficult and time-consuming, limiting label volume and quality. Thus, the findings primarily indicate the need to expand dataset size/diversity and refine annotation strategies, rather than reflecting insufficient model capacity. To ensure transparency and reproducibility, the BDR-iD dataset is publicly available in an open repository: https://github.com/carlossantos-iffar/BDR-iD-dataset.

CONCLUSION

This paper introduces the preliminary Brazilian Diabetic Retinopathy Images Dataset (BDR-iD), collected from a single ophthalmology clinic, and benchmarks its first release using deep learning for DR grading, lesion segmentation, and lesion detection. The results show that lesion-level tasks are substantially harder than DR grading and will require larger, more diverse sets of expert-annotated images. On the test set, RegNet Y 400MF achieved the best DR grading performance (overall accuracy 0.6667, average accuracy 0.4634, Kappa 0.5179). For lesion segmentation, R2U-Net produced the most balanced results, while YOLOv11 achieved the best detection performance (mAP 0.3460 on validation and 0.3810 on testing). Although these metrics are moderate, they are expected given the limited sample size and lesion distribution in BDR-iD v1 and should be interpreted as baseline results.

Future work will expand BDR-iD with data from additional Brazilian centers, increase expert-validated lesion annotations, and evaluate semi-supervised and self-supervised methods to exploit the large pool of unlabeled images. We will also assess classical supervised machine learning with handcrafted features as a potentially competitive, lower-cost alternative on larger and more diverse datasets, supporting robust benchmarking and real-world deployment in Brazilian healthcare settings.

ACKNOWLEDGEMENTS

This work is partly financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), Finance Code 001.

ETHICS APPROVAL

The research was registered on Plataforma Brasil and approved by the Research Ethics Committee (CEP) of the Federal University of Santa Maria (UFSM), CAAE No. 7723523.1.000.5346, Consolidated Opinion No. 5.959.406.

USE OF GENERATIVE AI TOOLS

We used generative AI tools only to assist with minor language editing during manuscript revision. All study design, data collection, data analysis, and scientific interpretation were conceived and performed by the authors, who take full responsibility for the content of this work.

REFERENCES

1. International Council of Ophthalmology. Updated 2017 ICO Guidelines for Diabetic Eye Care [Internet]. San Francisco (CA): International Council of Ophthalmology; 2017 [cited 2025 Jun 6]. Available from: https://icoph.org/eye-care-delivery/diabetic--eye-care/
2. Nayak J, Bhat PS, Acharya UR, Lim CM, Kagathi M. Automated identification of diabetic retinopathy stages using digital fundus images. J Med Syst. 2008;32(2):107–15.
3. Ting DSW, Cheung GCM, Wong TY. Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review. Clin Exp Ophthalmol. 2016;44(4):260–77.
4. Chakrabarti R, Harper CA, Keeffe JE. Diabetic retinopathy management guidelines. Expert Rev Ophthalmol. 2012;7(5):417–39. [cited 2025 Jun 6] Available from: https://doi.org/10.1586/eop.12.52
5. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [Internet]. 2015 [cited 2025 jun 6]. Available from: https://arxiv.org/abs/1409.1556
6. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition [Internet]. 2015 [citado 2025 jun 6]. Available from: https://arxiv.org/abs/1512.03385
7. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions [Internet]. 2014 [cited 2025 Jun 6]. Available from: https://arxiv.org/abs/1409.4842
8. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks [Internet]. 2018 [cited 2025 Jun 6]. Available from: https://arxiv.org/abs/1608.06993
9. Tan M, Le QV. EfficientNet: Rethinking model scaling for convolutional neural networks [Internet]. 2020 [citado 2025 jun 6]. Available from: https://arxiv.org/abs/1905.11946
10. Radosavovic I, Kosaraju RP, Girshick R, He K, Dollár P. Designing network design spaces [Internet]. 2020 [cited 2025 Jun 6]. Available from: https://arxiv.org/abs/2003.13678
11. Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size [Internet]. 2016 [cited 2025 Jun 6]. Available from: https://arxiv.org/abs/1602.07360
12. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin Transformer: Hierarchical vision transformer using shifted windows [Internet]. 2021 [cited 2025 Jun 6]. Available from: https://arxiv.org/abs/2103.14030
13. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Lect Notes Comput Sci. 2015;9351:234–41.
14. Oktay O, Schlemper J, Le Folgoc L, Lee M, Heinrich M, Misawa K, et al. Attention U-Net: Learning where to look for the pancreas. arXiv Preprint. 2018; arXiv:1804.03999.
15. Alom MZ, Hasan M, Yakopcic C, Taha TM, Asari VK. Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation. arXiv Preprint. 2018; arXiv:1802.06955.
16. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for the detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316(22):2402–10.
17. Ting DSW, Cheung CY, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA. 2017;318(22):2211–23.